Abstract

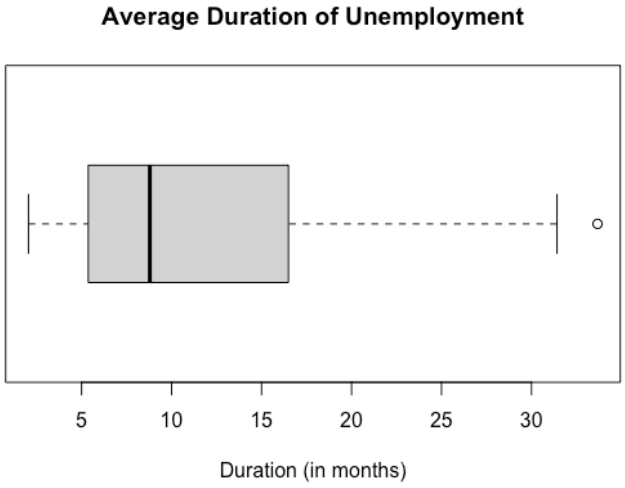
Understanding patterns of unemployment between age groups is important in informing policy intending to maximize employment in a country. If certain age groups experience longer durations of unemployment than other age groups, then that could be indicative of deeper problem within a country's labor structure. In this study, we looked at the duration of unemployment data collected from 9 OECD countries in 2017 and separated the data by age groups (15-19, 20-24, 25-54, 55-99). We found that there is a difference in mean duration of unemployment in the 9 OECD countries in 2017 between age groups 55-99 and 15-19 years as well as between 55-99 years and 20-24 years. There was no difference in mean duration of unemployment between any other age groups.

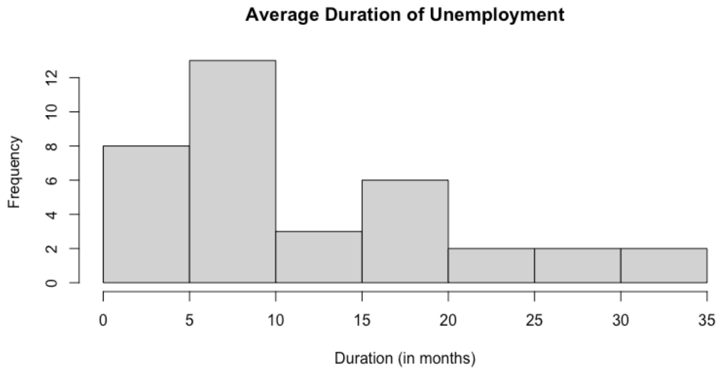
The Organization for Economic Co-operation and Development (OECD) is an intergovernmental organization where OECD-admitted governments come together to seek solutions to common macroeconomic issues (OECD, 2020). There are currently 37 in the OECD; to be admitted to the OECD, countries must adhere to two fundamental requirements: "(i) democratic societies committed to rule of law and protection of human rights; and (ii) open, transparent and free-market economies" (OECD, 2018).

The goal of this project is to see whether the duration of unemployment (in months) of OECD countries differs by age group. The data set looks at 9 different OECD countries (Australia, Canada, Czech Republic, Finland, France, Hungary, Norway, Slovak Republic, United States) and the duration of unemployment (in months) for each age group (15-19 years, 20-24 years, 25-54 years, 55-99 years) in those countries in 2017. The data set was found on the OECD.Stat which allows users to search for and extract data from across the OECD databases. The data was initially collected via survey by the governments of the aforementioned countries and then relayed to the OECD (OECD, 2020). The specific surveys each government created and their specific definitions of "unemployed" can be found on pages 13-20 in the labor force statistics reference guide for the OECD, which is cited in the bibliography.

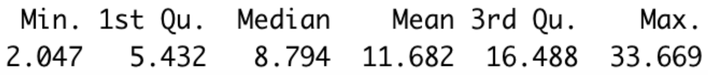
Data

To be clear, the numeric variable that will be explored via ANOVA is the mean duration of unemployment (in months) of the 9 OECD countries in 2017. The categorical variable is age group (in years) and the levels of age group are (15-19, 20-24, 25-54, 55-99). Looking at the raw data, the column "Value" is the duration of unemployment (in months) while "AGE" is the age group where ("1519" corresponds to 15-19 years age group, "2024" corresponds to 20-24 years age group, "2554" corresponds to 25-54 years age group, and "5599" corresponds to 55-99 years age group).



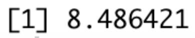


Summary statistics of overall data:



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.047   5.432   8.794  11.682  16.488  33.669
```

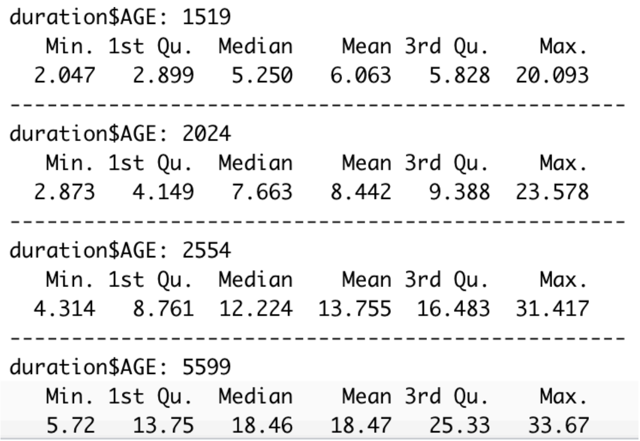
Standard Deviation:



```
[1] 8.486421
```

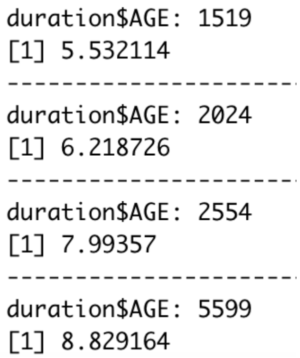
Looking at the summary statistics, boxplot, and histogram, the overall data seems to be skewed right with one outlier at 33.669 months. The median looks to be around 8 months and this is confirmed by the summary statistics at 8.794 months. The mean is a little bit more than the median at 11.682 months, which makes sense because the data set is skewed right. The spread is from 2.047 months to 33.669 months.

Summary Statistics by Age Group:

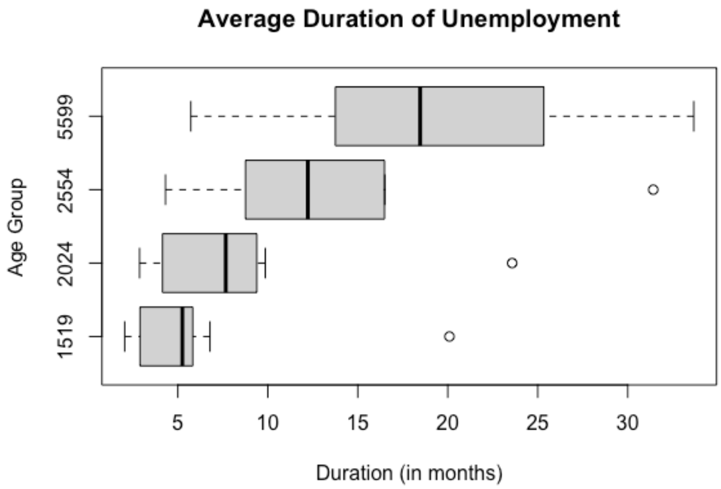


```
duration$AGE: 1519
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.047   2.899   5.250   6.063   5.828  20.093 
------------------------------------------------- 
duration$AGE: 2024
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.873   4.149   7.663   8.442   9.388  23.578 
------------------------------------------------- 
duration$AGE: 2554
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.314   8.761  12.224  13.755  16.483  31.417 
------------------------------------------------- 
duration$AGE: 5599
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5.72   13.75   18.46   18.47   25.33   33.67 
```

Standard Deviation:



```
duration$AGE: 1519
[1] 5.532114
----------------------
duration$AGE: 2024
[1] 6.218726
----------------------
duration$AGE: 2554
[1] 7.99357
----------------------
duration$AGE: 5599
[1] 8.829164
```



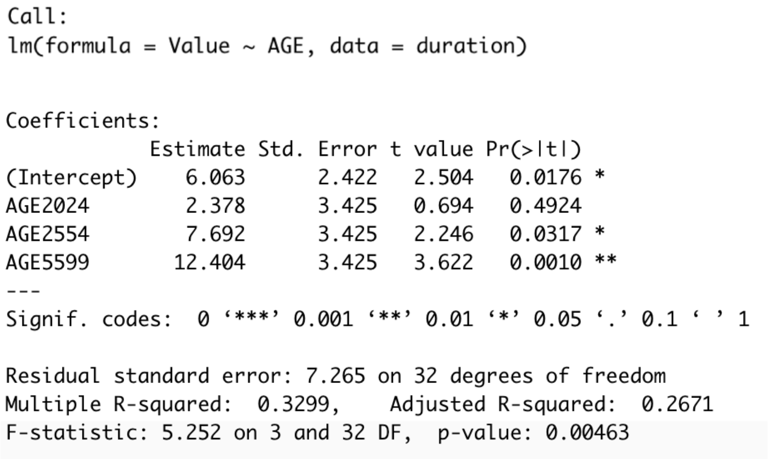
The shapes data of the age groups differ. The 55-99 age group looks to be normally distributed while the 25-54 age group looks to be skewed right. There are high outliers in each of the distributions of the age groups except for the 55-99 age group, and because the sample sizes for each of those age groups (N=9) is relatively small, the outliers could affect the statistical findings. The means for each of the age groups from youngest to oldest is 6.063 months, 8.442 months, 13.755 months, and 18.47 months. The spread of the data for each of the age groups also seem to differ substantially. Notably, the minimum duration of unemployment for the 55-99 age group is around the center for the duration of unemployment for the 15-19 age group while the maximum duration for that oldest group encompasses all of the outliers of the younger groups.

Looking at the summary statistics and side-by-side boxplot, it does look like there is a difference in the mean duration of unemployment in the 9 OECD countries by age group. The mean duration of unemployment for 15-19 year olds is 6.063 months while for the 55-99 age group it is 18.47 months, which is triple what it is for the former age group. However, the 20-24 age group mean is 8.442 months, which does not seem to differ too much from the 15-19 age group. The 25-55 age group mean is 13.755 months which is more than double the 15-19 age group.

Summary of Statistical Findings

In order to see if there were statistical differences in mean duration of unemployment among the four different age groups, we constructed an ANOVA model with duration of unemployment as the numerical response variable and age groups as the categorical explanatory variable.

The result of the ANOVA model is as follows:



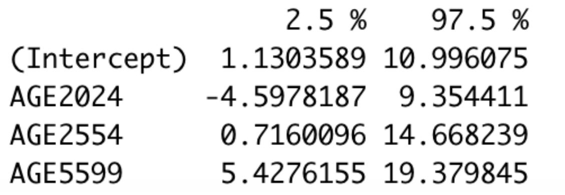
```
Call:
lm(formula = Value ~ AGE, data = duration)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.063      2.422   2.504   0.0176 *
AGE2024        2.378      3.425   0.694   0.4924
AGE2554        7.692      3.425   2.246   0.0317 *
AGE5599       12.404      3.425   3.622   0.0010 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.265 on 32 degrees of freedom
Multiple R-squared:  0.3299,    Adjusted R-squared:  0.2671
F-statistic: 5.252 on 3 and 32 DF,  p-value: 0.00463
```

The fitted model equal is as follows:

$$\mu[Value|AGE] = 6.063 + 2.378 * I[AGE = 2024] + 7.692 * I[AGE = 2554] + 12.404 * I[AGE = 5599$$

Level	**Est for population mean**	**Null hypothesis**	**Statistically significant?**
0 - 1519	6.063	The population mean duration of unemployment of 15-19 year olds is 0.	Yes – P-value = 0.0176, meaning there is sufficient evidence to reject the null hypothesis and believe that the population mean duration of unemployment of 15-19 year olds is not 0.
1 - 2024	6.063+2.378 = 8.441	The difference in population mean duration of unemployment of 15-19 year olds and 20-24 year olds is 0.	No – P-value = 0.4924, meaning there is not sufficient evidence to reject the null hypothesis and believe that the difference in population mean duration of unemployment between 15-19 year olds and 20-24 year olds is not 0.
2 - 2554	6.063+7.692=13.755	The difference in population mean duration of unemployment of 15-19 year olds and 25-54 year olds is 0	Yes – P-value = 0.0317, meaning there is sufficient evidence to reject the null hypothesis and believe that the difference in population mean duration of unemployment between 15-19 year olds and 25-54 year olds is not 0.
3 - 5599	6.063+12.404=18.467	The difference in population mean duration of unemployment of 15-19 year olds and 55-99 year olds is 0.	Yes – P-value = 0.0010, meaning there is sufficient evidence to reject the null hypothesis and believe that the difference in population mean duration of unemployment between 15-19 year olds and 55-99 year olds is not 0.

A 95% confidence interval for each coefficient as follows:



```
                2.5 %    97.5 %
(Intercept)  1.1303589 10.996075
AGE2024     -4.5978187  9.354411
AGE2554      0.7160096 14.668239
AGE5599      5.4276155 19.379845
```

We are 95% confident that the true mean duration of unemployment for 15-19 year olds from the select 9 OECD countries is between 1.1303589 and 10.996075 months.

We are 95% confident that the true mean duration of unemployment for 20-24 year olds from the select 9 OECD countries is between -4.5978181 and 9.354411 months.

We are 95% confident that the true mean duration of unemployment for 25-54 year olds from the select 9 OECD countries is between 0.7160096 and 14.668238 months.

We are 95% confident that the true mean duration of unemployment for 55-99 year olds from the select 9 OECD countries is between 5.4276155 and 19.379845 months.

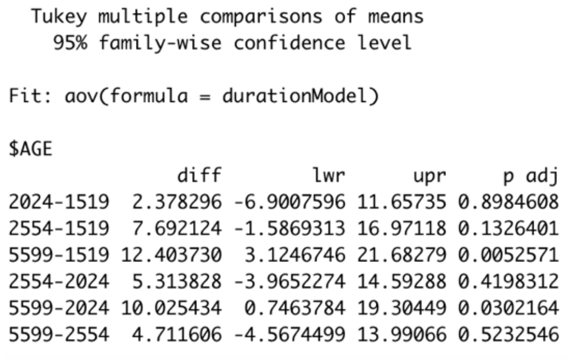
Discussion of the F-test is as follows:

H0: all mean durations of unemployment are the same (for each age group)

Ha: the mean duration of unemployment is different for at least one age group

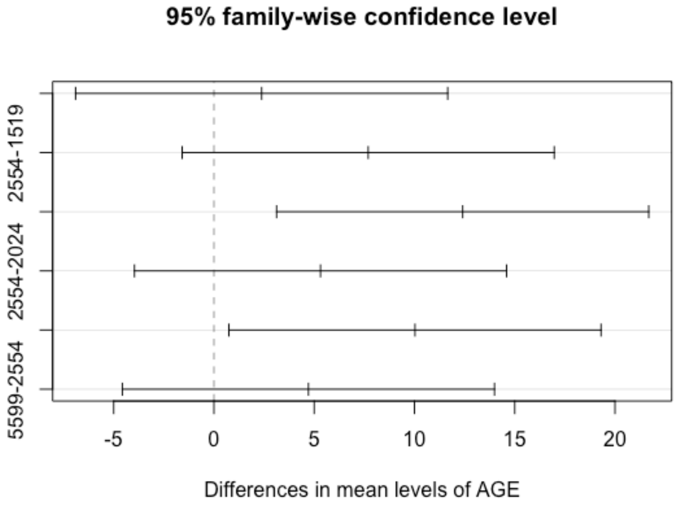
Because the p-value is 0.00463 and the F-statistic is 5.252 this means that we reject the null hypothesis and believe at least one mean duration of unemployment is different for at least one age group, meaning that the ANOVA model is useful.

A Tukey post-hoc test was performed see differences in means between two age groups. The null hypothesis of the pairwise comparisons done by the Tukey post-hoc test is that there is no difference in the means. The alternative hypothesis is that there is a difference between the two groups. The result of the Tukey post-hoc test is displayed below:



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = durationModel)

$AGE
               diff        lwr      upr     p adj
2024-1519  2.378296 -6.9007596 11.65735 0.8984608
2554-1519  7.692124 -1.5869313 16.97118 0.1326401
5599-1519 12.403730  3.1246746 21.68279 0.0052571
2554-2024  5.313828 -3.9652274 14.59288 0.4198312
5599-2024 10.025434  0.7463784 19.30449 0.0302164
5599-2554  4.711606 -4.5674499 13.99066 0.5232546
```

The Tukey HSD plot visualizing all the intervals of the compared age groups.



2024-1519: Because the p-value is 0.8984608 and the confidence interval contains 0, we fail to reject the null hypothesis that there is a difference in the duration of unemployment between these two groups.

2554-1519: Because the p-value is 0.1326401 and the confidence interval contains 0, we fail to reject the null hypothesis that there is a difference in the duration of unemployment between these two groups.

5599-1519: Because the p-value is 0.0052571 and the confidence interval does not contain 0, we reject the null hypothesis and believe that there is a difference in the duration of unemployment between these two groups.

2554-2024: Because the p-value is 0.4198312 and the confidence interval contains 0, we fail to reject the null hypothesis that there is a difference in the duration of unemployment between these two groups.

5599-2024: Because the p-value is 0.0302164 and the confidence interval does not contain 0, we reject the null hypothesis and believe that there is a difference in the duration of unemployment between these two groups.

5599-2554: Because the p-value is 0.5232546 and the confidence interval contains 0, we fail to reject the null hypothesis that there is a difference in the duration of unemployment between these two groups

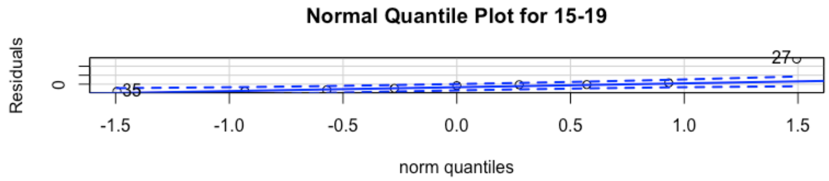
The graphs, statistics, and ANOVA model suggest there is a difference in the mean duration of unemployment in 9 OECD countries in 2017 by age group, but only a particular few age groups. While there was a statistically significant difference in mean duration of unemployment in 9 OECD countries in 2017 between the 55-99 age group and the 15-19 and 20-24 age groups, there is not statistically significant difference in the mean duration of unemployment in 9 OECD countries in 2017 between any other age groups reported. From the outset, by looking at the exploratory data section and side-by-side boxplot, the biggest differences in center were between the 55-99 age group and the two youngest age groups. The Tukey HSD plot and post-hoc test is helpful in the confirmation of this belief. Looking at the Tukey post-hoc test, the only age group comparisons that reported a p-value of less than $\alpha = 0.05$ that allowed us to reject the null hypothesis were 55-99 year olds vs 15-19 year olds and 55-99 year olds vs 20-24 year olds. While the comparison between 15-19 year olds and 25-54 year olds was statistically significant in the ANOVA model, because the Tukey post-hoc suggested that that comparison was not statistically significant, we will not consider a statistically significant difference between the mean durations of unemployment in 9 OECD in 2017 between those two age groups.

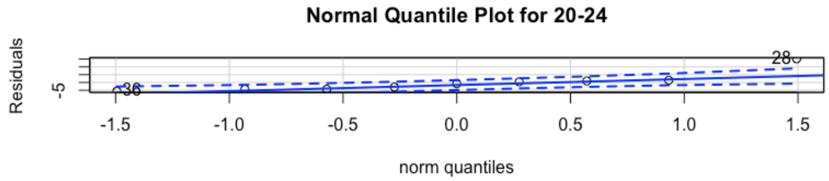
Checking the Sampling Variability Assumptions:

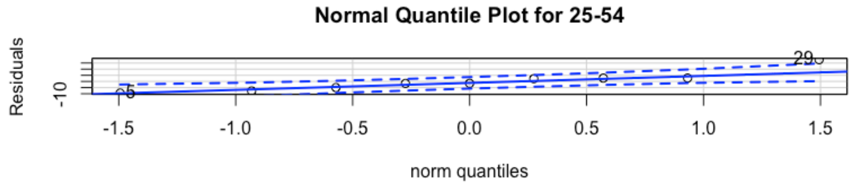
The ANOVA model used has two sampling variability assumptions (SVAs)

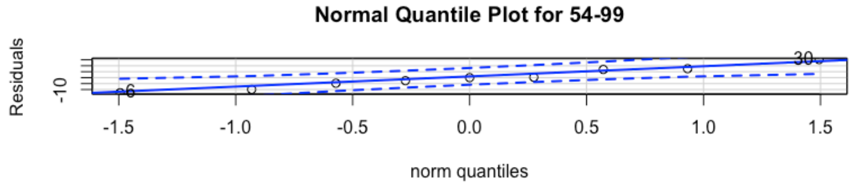
1. The errors (residuals) are normally distributed
2. The errors should have equal variances

In order to test the first SVA, we constructed normal quantile plots for the residuals of each of the age groups in the model.

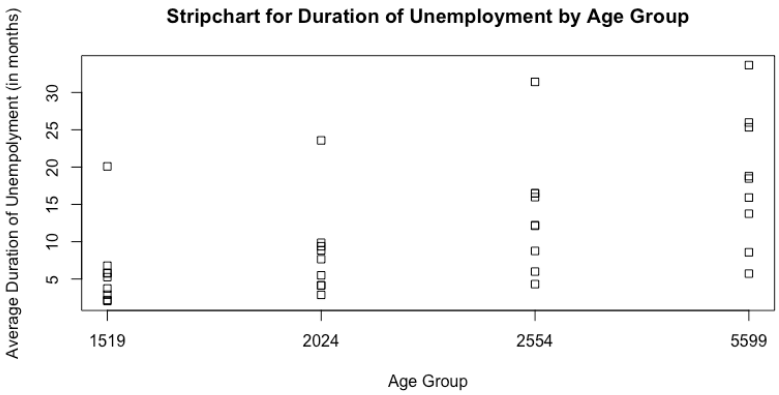


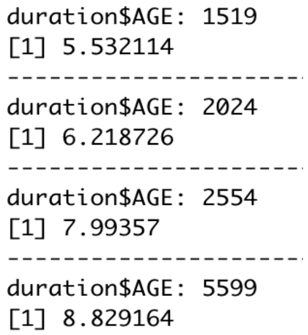






From the plots shown above, it is seen that all of the points lie within the dashed lines for the 54-99 age group. However, for the other plots, there is one point that is consistently outside the dashed lines. This point corresponds to the outliers previously noted in the data exploration section. Since the sample size is small (N=9), the outlying point in each of those plots is strong evidence to believe that the data in regard to duration of unemployment in the age groups 15-19, 20-24, and 25-54 is not normal. This deviation from normality could affect the accuracy of the ANOVA test.





```
duration$AGE: 1519
[1] 5.532114
---------------------
duration$AGE: 2024
[1] 6.218726
---------------------
duration$AGE: 2554
[1] 7.99357
---------------------
duration$AGE: 5599
[1] 8.829164
```

The second SVA is that the age groups have equal variances. By looking at the strip chart above, it can be seen that the spread of the data is more or less similar to each other. From the descriptive statistics, the highest standard deviation (8.829164) is not twice as great as the lowest standard deviation (5.532114). Therefore, there is no evidence to believe that the variances differ greatly from each other, so it can be concluded that the second SVA is satisfied.

Scope of Inference

Can we infer cause and effect? i.e. Does being in a certain age group make it where people are unemployed for longer (on average)? The study did not control for any factors that could have influenced the results of the data. The study used observationally reported data. If a randomized comparative experiment were to be done in order to infer cause and effect, we would have to randomly assign some subjects to be in a certain age group, but clearly that is impossible.

Can we generalize these results to a larger population? i.e. Can it be said that all unemployed people have different lengths of unemployment (on average) based on their age group? The subjects of the study were from 9 OECD countries and they were not an SRS even from that population. It is reasonable to believe, however, that the individuals from the sample are representative of all unemployed people in those 9 OECD countries in 2017. We might be able to generalize these findings to age groups of other OECD countries in 2017, based on the aforementioned admittance requirements to OECD. To other countries beyond that, generalizing would be a lot more speculative.

Bibliography

OECD. (2018). *About the OECD - OECD*. Oecd.Org. https://www.oecd.org/about/

OECD. (2017). *Average duration of unemployment*. Stats.Oecd.Org. https://stats.oecd.org/Index.aspx?DataSetCode=AVD_DUR

OECD. (2020). *LABOUR FORCE STATISTICS IN OECD COUNTRIES: SOURCES, COVERAGE AND DEFINITIONS* (pp. 13–20). http://www.oecd.org/els/emp/LFS%20Definitions%20-%20Tables.pdf