

# Data Analytics II: Project Part A

57

Gabriela Siren, Didrik Gentili, Anna Gao, Ludwig Fredriksson  
25601, 25550, 25811, 25632

2023-01-29

## A.1 - Explorative Data Analysis (EDA)

Table 1: Summary table for all variables

	Age	Income	Kids		Gender	OwnsHome	Subscribes
Mean	40.74	51901.34	1.32	Female or False	0.523	0.517	0.867
Stan. Deviaton	12.84	20160.88	1.38	Male or True	0.477	0.483	0.133
Minimum	17.00	11165.00	0.00				
Q1 .25%	33.00	39906.75	0.00				
Median	39.00	52574.00	1.00				
Q3 .75%	48.25	64865.50	2.00				
Maximum	77.00	138959.00	7.00				

As shown by Table 1, ages range between 17 to 77 with a median of 39. The mean age is 40.74 which is slightly larger than the median. Additionally, the majority of the ages are in between 33 and 48.25.

The income has a very large range, from roughly 11 000 to 139 000. The median, 53000 and mean, 52000 are closer to the minimum than the maximum meaning the data is skewed towards the lower values. The 3rd quartile is also very close to the median/mean meaning that there are some potential extreme values.

The kids are also skewed towards lower values. The range goes from 0 to 7 but the mean is 1.32 while the median is 1. This all gives us an accurate picture of the distribution of the data.

On the other hand, there is a relatively even proportion between two of the categorical variables, excluding the subscribers. While there is roughly 4 percent difference between males and females, they are both around 50%. The same applies for people owning homes. However, the proportion of subscribers is not distributed evenly as more than 85% of the sample does not subscribe. While this tells that the majority do not subscribe and that the sample has an even proportion of categories, this does not give us specific categorical values, for example, whether more Females or Males subscribe.

Table 2: Summary table for age

Segment	Mean	StDev	Minimum	Maximum	Q1	Median	Q3
Moving up	36.01	4.24	27	47	33	35	38.00
Suburb mix	39.29	5.14	27	53	36	39	42.25
Travelers	57.71	7.75	43	77	52	58	62.25
Urban hip	23.08	2.21	17	28	22	23	24.00

Table 3: Summary table for income

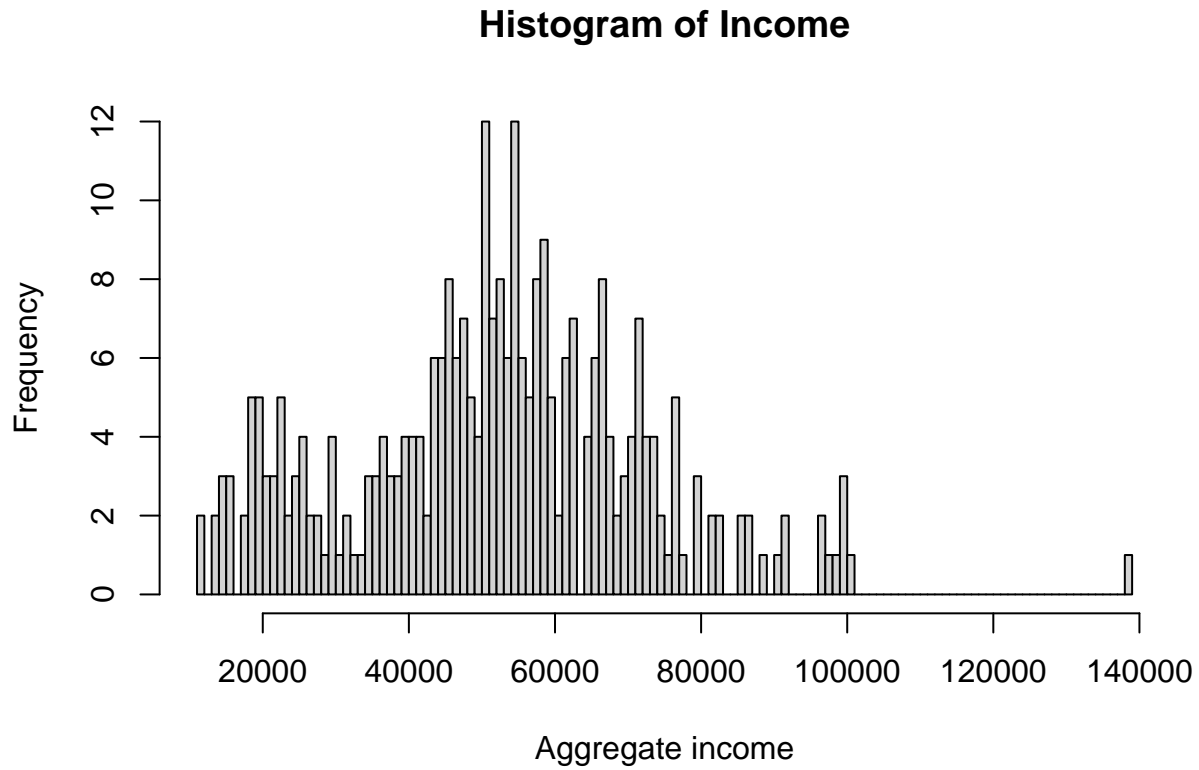
Segment	Mean	StDev	Minimum	Maximum	Q1	Median	Q3
Moving up	50613.60	8805.88	34975	73554	45391.00	50746.5	55835.75
Suburb mix	55444.01	12391.52	26521	82911	47162.75	55095.0	65273.00
Travelers	67832.45	19695.61	29699	138959	53802.50	66614.5	80348.00
Urban hip	21129.08	5175.90	11165	34106	18201.25	20850.0	24536.75

Table 4: Summary table for children

Segment	Mean	StDev	Minimum	Maximum	Q1	Median	Q3
Moving up	1.69	1.26	0	5	1	1	2
Suburb mix	1.90	1.49	0	7	1	2	3
Travelers	0.27	0.45	0	1	0	0	1
Urban hip	1.30	1.39	0	5	0	1	2

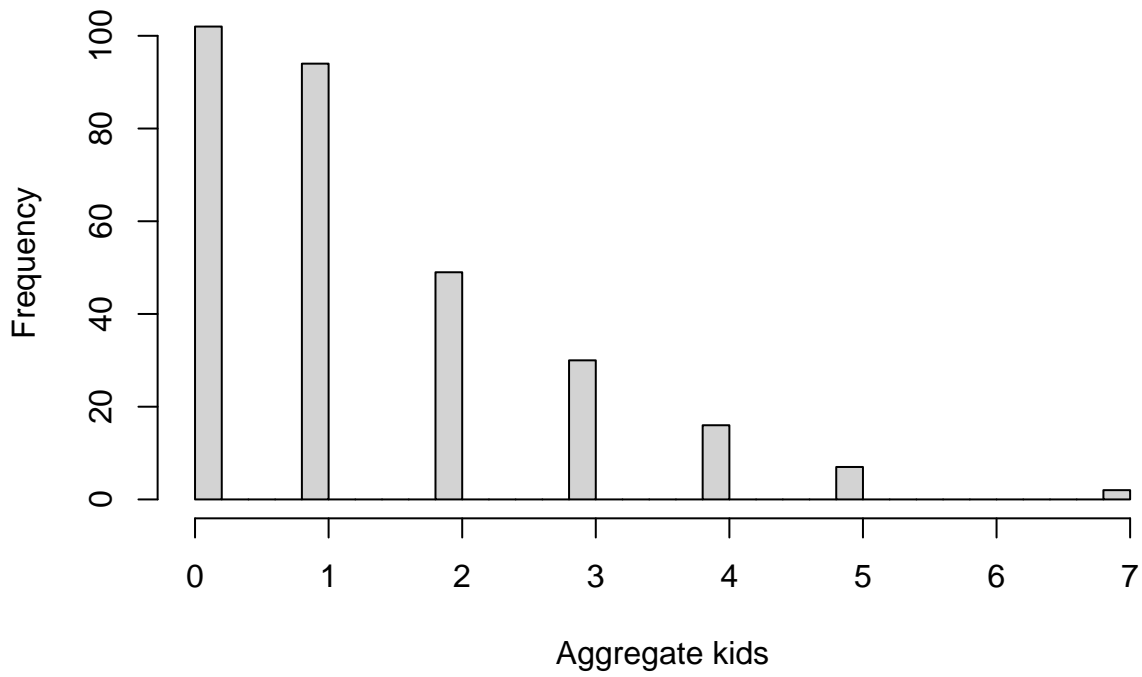


The histogram is very centered around the mean while the median is a bit greater than where the majority lie. Additionally, there seems to be an observation above all the other ones at 77 so it could potentially be an outlier. The distribution could have been normally distributed but when increasing the number of breaks, it was revealed that there are almost two modes. However, it is still possible to approximate this distribution to a normal one.



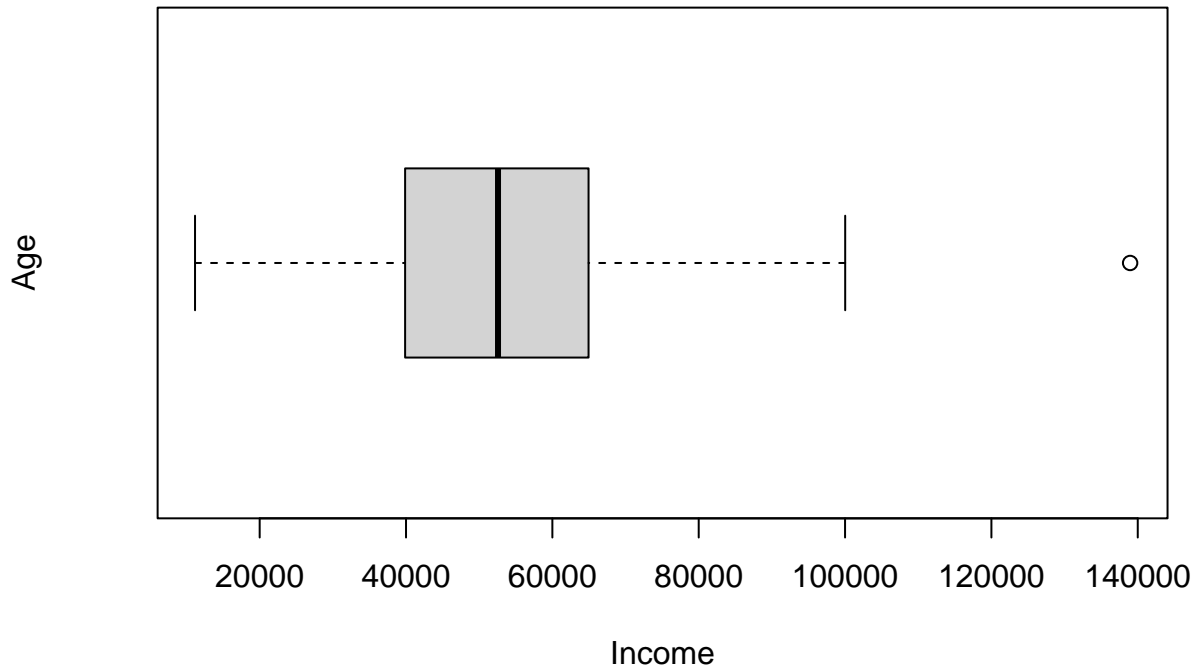
The income histogram is also centered around its mean and median. There is an observation at roughly 139 000 which is definitely an outlier in this case. Excluding the outlier, the above sample can be approximated to a normal distribution. However, there is a small peak around 20 000 which might contradict that approximation.

### Histogram of Kids



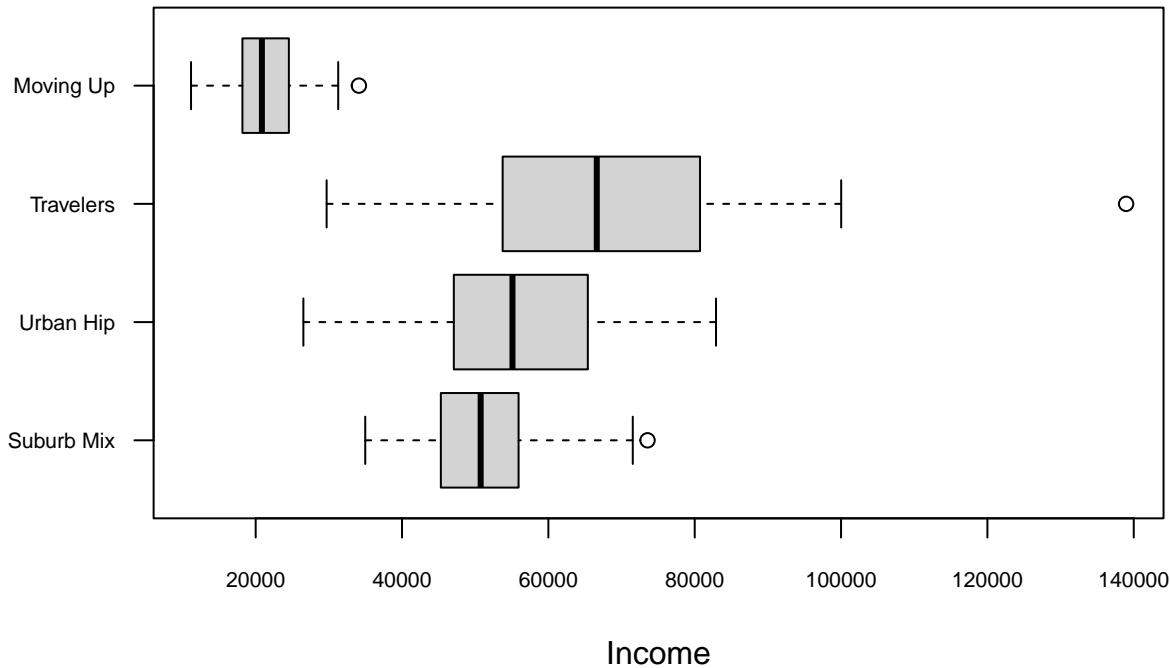
The histogram of Kids is centered around the minimum rather than median or the mean and this could be because of the outlier at 7. The median in this case is more central than the mean as a result of this outlier.

### Aggregate Income boxplot



The Aggregate Income boxplots supports the claims made from the histogram. While there is quite a even distribution centered around the median, there is an extreme value at around 139 000. The quartiles, and the maximum and minimum are also more or less evenly distributed.

## Boxplots for Segmented Income



| As shown by the boxplot above, there is a significant difference between the data from different segments. For example, the median moving up income is around 20 000 while for Travelers it is 65 000. Urban hip and Suburb mix have a slightly lower median income at around 52 000 and 50 000 respectively. Furthermore, while the Travelers segment has a far outlier, it does not affect the median so the current boxplot is accurate.

| Additionally, there is also a much lower minimum and maximum value for Moving Up compared to the rest. The Moving Up and Urban Hip segment is also not skewed compared to the other two. Travelers is slightly skewed towards the left while Suburban mix is skewed to the right.

Do the variables Income and Age appear to be (approximately) normally distributed?

The Income and Age variables appear to be approximately normally distributed, even if there appears to be two “modes” in the Age variable, as mentioned above.

Can you think of a suitable distribution for the variable children? | The Children variable is not approximately normally distributed as it appears rather like a Chi-squared distribution. The distribution is very skewed to the right, with the mode furthest to the left.

## A.2 - Confidence Intervals

```
## [1] "aggregate lever income confidence interval at 90% confidence"
```

```
## [1] 51901.34
```

```
## [1] 49980.8
```

```
## [1] 53821.89
```

[your comments here]

[your comments here]

Are the underlying assumptions for calculating this interval met in this data?

### A.3 - Confidence Intervals and the Sample Size

```
## [1] 1690.965
```

```
## [1] 6763.859
```

[your comments here]

```
## [1] 6763.859
```

How should we adjust the confidence level as the size of the sample increases? Why?

### A.4 - Confidence Intervals: Comparison of Population Means

To determine the confidence interval for the difference of population means, we used the Welch's t-test on 80 samples of travelers and 50 samples from the urban hip segment. We assumed that observations in each sample were independently and randomly chosen, and that the populations follow normal distributions. The results showed that the 90% confidence interval for the difference of the mean income ( $\bar{x}_{travelers} - \bar{x}_{Urbanhip}$ ) was (42849, 50557), thus indicating a significant difference in the mean income between the two segments with a high degree of confidence.