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
Mean	40.74	51901.34	1.32
Stan. Deviaton	12.84	20160.88	1.38
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

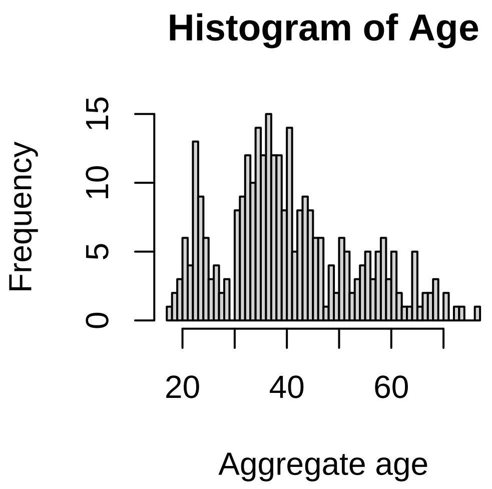
	Gender	OwnsHome	Subscribes
Female or False	0.523	0.517	0.867
Male or True	0.477	0.483	0.133

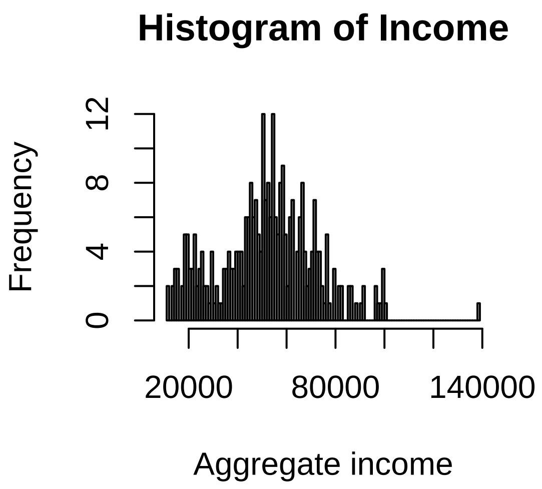
As shown by Table 1, ages range between 17 to 77 with a median of 39. The mean age is 40.74 which is slightly larger than the median. Additionally, the majority of the ages are in between 33 and 48.25.

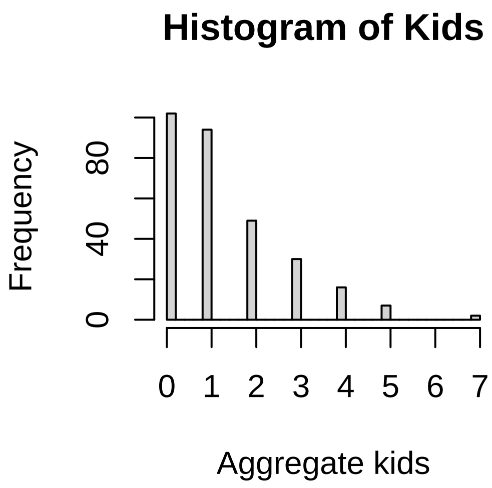
The income has a very large range, from roughly 11000 to 139000. The median, 53000 and mean, 52000 are closer to the minimum than the maximum meaning the data is skewed towards the lower values. The 3rd quartile is also very close to the median/mean meaning that there are some potential extreme values.

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. For example, While there is roughly 4 percent difference between males and females, they are both around 50. 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.



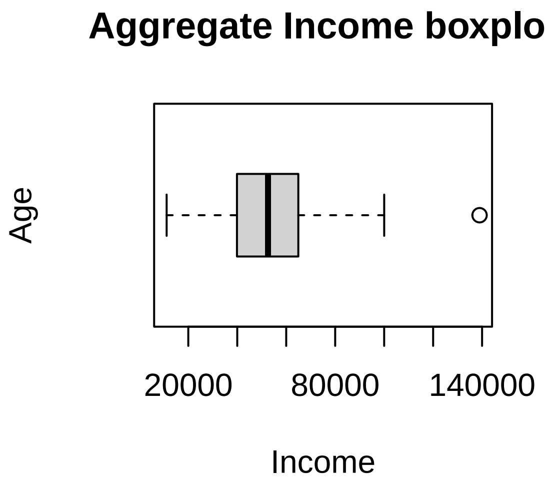


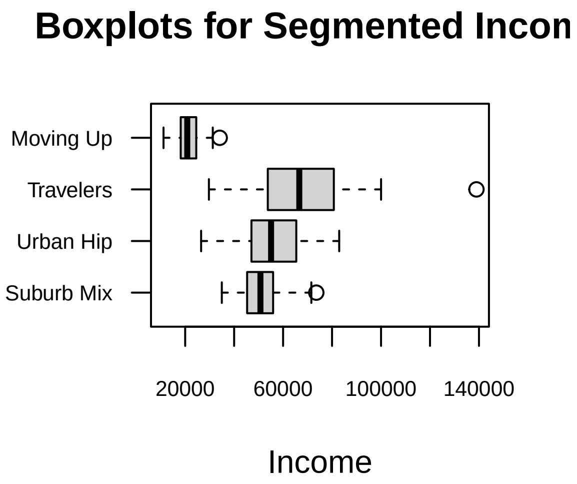


The histogram is very centered around the mean while the median is a bit greater than where the mode is. Additionally, there seems to be an observation at 77 which it could potentially be an outlier.

The income histogram is also centered around its mean and median. There is an observation at roughly 139000 which is definitely an outlier in this case.

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.





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 139000. The quartiles, and the maximum and minimum are also more or less evenly distributed.

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 20000 while for Travelers it is 65000. Urban hip and Suburb mix have a slightly lower median income at around 52000 and 50000 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? Can you think of a suitable distribution for the variable children?

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.

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

Table 2: T-test table for Aggregate Income, at a 90% C.I for the population mean

	Estimate	Statistic	P.Value	DF	LCL	UCL	ME
Aggregate	51901.34	44.59	0	299	49980.8	53821.89	1920.54

Table 3: T-test table for Segmented Income, at a 90% C.I for the population mean

Segments	Estimate	Statistic	P.Value	DF	LCL	UCL	ME
Travellers	67832.45	30.80	0	79	64167.44	71497.46	3665.01
Urban Hip	21129.08	28.87	0	49	19901.87	22356.29	1227.21
Suburb Mix	55444.01	44.74	0	99	53386.53	57501.49	2057.48
Moving Up	50613.60	48.09	0	69	48858.82	52368.38	1754.78

Table 4: T-test table for Income by Gender, at a 90% C.I for the population mean

Gender	Estimate	Statistic	P.Value	DF	LCL	UCL	ME
Male	49983.66	27.04	0	142	46923.64	53043.69	3060.03
Female	53648.02	37.12	0	156	51256.47	56039.57	2391.55

The population mean for income at an aggregate level has a confidence interval with a Lower Confidence Limit (LCL) of 49980.80 and Upper Confidence Limit (UCL) of 53821.89. The population mean lies within this interval with 90% confidence.

talk about segment

talk about gender

A.2.2

[your comments here]

Table 5: 90% C.I for the population proportion of Subscribers

	Estimate	Statistic	P.Value	DF	LCL	UCL	ME
Subscribers	0.13	6.78	0	299	0.1	0.17	0.03

Table 6: Proportion per Segment

	Proportion
Moving Up	0.17
Suburb Mix	0.09
Travelers	0.06
Urban Hip	0.28

Table 7: 90% C.I for the population proportion of Subscribers in the Urban Hip

Segment	Estimate	Statistic	P.Value	DF	LCL	UCL	ME
Urban Hip	0.28	4.37	0	49	0.17	0.39	0.11

A.2.3

[your comments here]

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

A.3 - Confidence Intervals and the Sample Size

Table 8: Sample size to obtain a certain ME

2% ME	1% ME
1690.96	6763.86

It may not be worth the cost since the sample size would have to increase by a lot in order lower the ME only a little bit. In this case from 1691 to 6764 (around 4 times) in order to only decrease the ME from 2% to 1%.

Table 9: 90% C.I for population proportion of Subscrbers with 10000 samples

	Estimate	Statistic	P.Value	DF	Conf.Low	Conf.High	ME
Aggregate	0.1333333	5376.311	0	1	0.1277919	0.1390749	0.0057416

[your comments here]

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 (8.8 Newbold p.317). 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.