
Background Information

This assignment is based on a dataset titled as 'BUSS613 Assignment Dataset'. The dataset provides demographic and household information for 5000 customers of a telecom service provider.

Annexures

Important information about the variables is presented in Annexure 1 through Annexure 4. This information will help complete the assignment.

A brief description of the annexures is presented below.

- Annexure 1.** This annexure provides information about the variable names (i.e., column heading) and their labels/ description.
- Annexure 2.** This annexure provides information about the categorical/ ordinal variables. It presents the labels/ descriptions of the codes of the categories within each categorical/ ordinal variable.
- Annexure 3.** This annexure provides a histogram and boxplot for each continuous variable.
- Annexure 4.** This annexure provides the distribution of frequencies for the categorical/ ordinal variables.

Assignment Tasks

For this assignment, you are required to perform the following tasks:

1. Understand the assignment dataset by going through the information given in the annexures and the excel spreadsheet.
2. Use the information given in the annexures to identify the potential data quality issues in the variables. Suggest a mechanism to deal with these data quality issues.
3. Identify one appropriate dependent variable that you would want to predict. Your dependent variable should be either ordinal or categorical. Provide the reasons for your choice of the dependent variable.
4. From the remaining variables, shortlist independent variables (maximum ten variables) for the prediction of your dependent variable.
5. Suggest one analytical technique that you will use for predicting the dependent variable. What will be your major considerations while using this technique? Justify your reasons for the choice of the analytical technique.

Assignment Formatting & Submission

The document (a word file) should not be more than 2000 words in length (font Times New Roman size 12; 1.5 line spacing; justified). Page margins are to be 2.5cm all around. The file should be submitted via Turnitin on the blackboard. The blackboard link for submission is *Assignment Submission*.

The assignment questions should be answered in the following template:

No.	Question	Answer
1.	Use the information given in the annexures to identify the potential data quality issues in the variables.	
2.	Suggest a mechanism to deal with these data quality issues.	
3.	Identify one appropriate dependent variable that you would want to predict. Your dependent variable should be either ordinal or categorical. Provide the reasons for your choice of the dependent variable.	
4.	From the remaining variables, shortlist independent variables (maximum ten variables) for the prediction of your dependent variable.	
5.	Suggest one analytical technique that you will use for predicting the dependent variable. What will be your major considerations while using this technique?	
6.	Justify your reasons for the choice of the analytical technique.	

Annexure 1: Variable Labels

S. No	Variable	Label
1.	custid	Customer ID
2.	region	Geographic indicator
3.	townsize	Size of hometown
4.	gender	Gender
5.	age	Age in years
6.	agecat	Age category
7.	ed	Years of education
8.	edcat	Level of education
9.	jobcat	Job category
10.	employ	Years with current employer
11.	empcat	Years with current employer
12.	retire	Retired
13.	income	Household income in thousands
14.	inccat	Income category in thousands
15.	debtinc	Debt to income ratio (x100)
16.	creddebt	Credit card debt in thousands
17.	othdebt	Other debt in thousands
18.	jobsat	Job satisfaction
19.	marital	Marital status
20.	reside	Number of people in the household
21.	pets	Number of pets owned
22.	homeown	Home ownership
23.	hometype	Building type
24.	address	Years at current address
25.	addresscat	Years at current address
26.	cars	Number of cars owned/leased
27.	commute	Primary commute transportation
28.	commutecat	Commute category
29.	commutetime	Commute time in minutes
30.	polview	Political outlook
31.	polparty	Political party membership
32.	polcontrib	Political contributions
33.	vote	Voted in the last election
34.	card	Primary credit card
35.	cardtype	Designation of primary credit card
36.	cardbenefit	Benefit program for primary credit card
37.	cardfee	The annual fee for primary credit card
38.	cardtenure	Years held the primary credit card
39.	cardtenurecat	Years held the primary credit card
40.	owntv	Owns TV
41.	ownvcr	Owns VCR
42.	owndvd	Owns DVD player
43.	owncd	Owns stereo/CD player
44.	ownpda	Owns PDA
45.	ownpc	Owns computer
46.	ownipod	Owns a portable digital audio player
47.	owngame	Owns a gaming system
48.	ownfax	Owns fax machine
49.	news	Newspaper subscription

Annexure 2: Variables and their categories

Value		Label
region	1	Zone 1
	2	Zone 2
	3	Zone 3
	4	Zone 4
	5	Zone 5
townsize	1	> 250,000
	2	50,000-249,999
	3	10,000-49,999
	4	2,500-9,999
	5	< 2,500
gender	0	Male
	1	Female
agecat	1	<18
	2	18-24
	3	25-34
	4	35-49
	5	50-64
	6	>65
	9	No response
edcat	1	Did not complete high school
	2	High school degree
	3	Some college
	4	College degree
	5	Post-undergraduate degree
jobcat	1	Managerial and Professional
	2	Sales and Office
	3	Service
	4	Agricultural and Natural Resources
	5	Precision Production, Craft, Repair
	6	Operation, Fabrication, General Labor
empcat	1	Less than 2
	2	2 to 5
	3	6 to 10
	4	11 to 15
	5	More than 15
retire	0	No
	1	Yes
inccat	1	Under \$25
	2	\$25 - \$49
	3	\$50 - \$74
	4	\$75 - \$124
	5	\$125+
jobsat	1	Highly dissatisfied
	2	Somewhat dissatisfied
	3	Neutral
	4	Somewhat satisfied
	5	Highly satisfied
marital	0	Unmarried
	1	Married
homeown	0	Rent
	1	Own
hometype	1	Single-family
	2	Multiple-Family

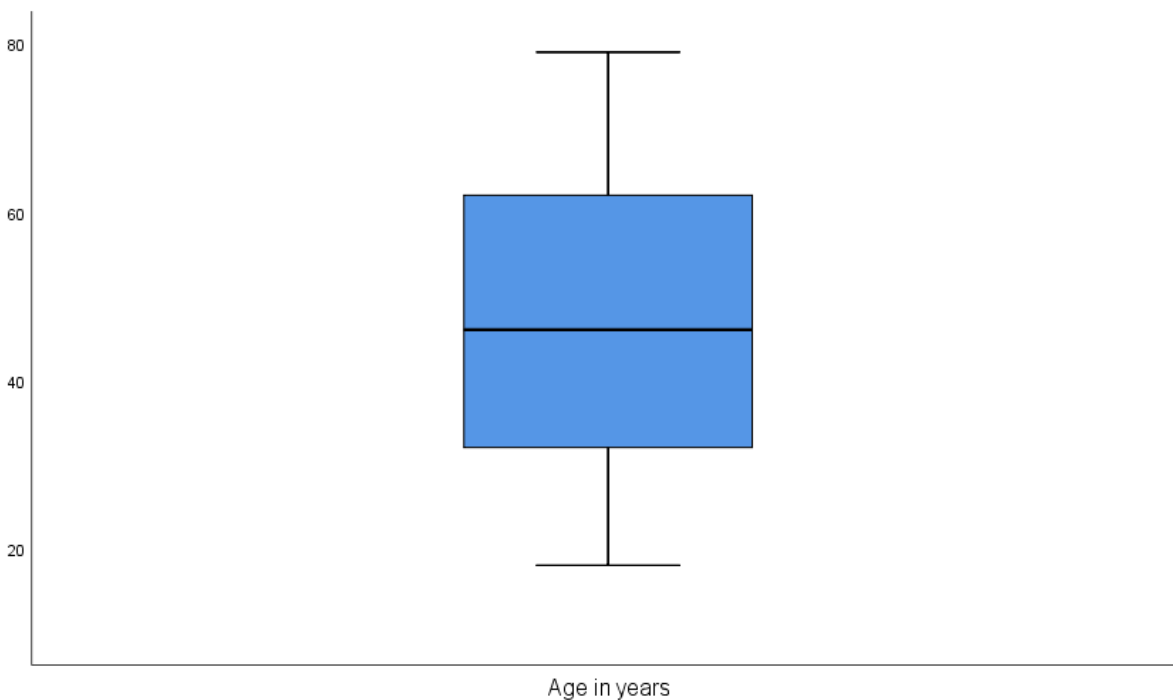
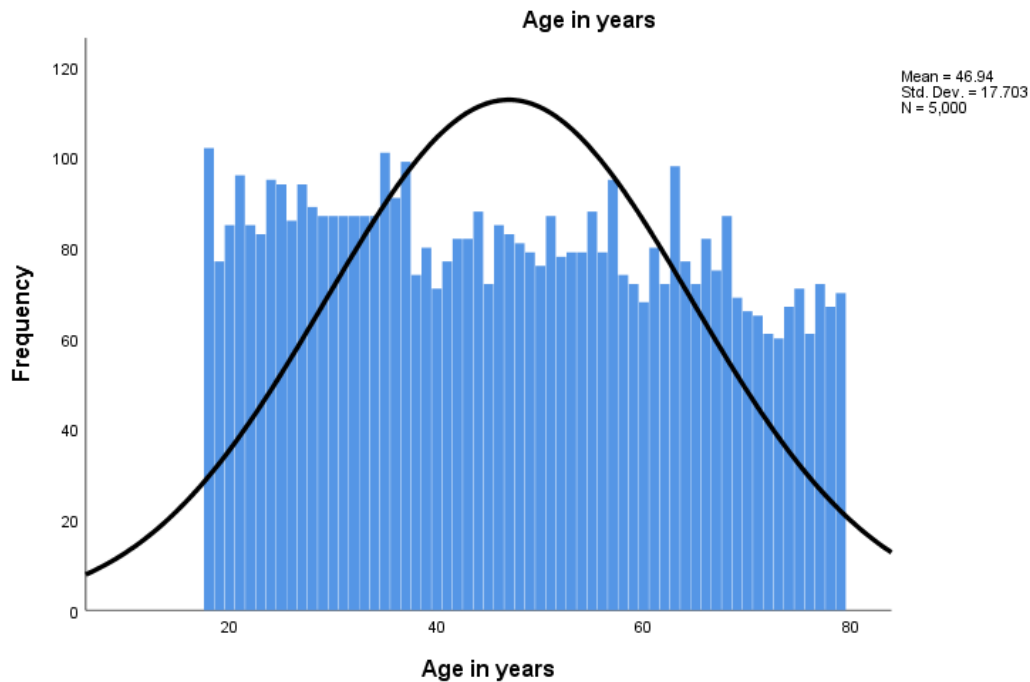
	3	Condominium/Townhouse
	4	Mobile Home
addresscat	1	Less than 3
	2	4 to 7
	3	8 to 15
	4	16 to 25
	5	More than 25
commute	1	Car
	2	Motorcycle
	3	Carpool
	4	Bus
	5	Train/Subway
	6	Other public transit
	7	Bicycle
	8	Walk
	9	Other non-motorized transit
	10	Telecommute
commutecat	1	Single occupancy
	2	Multiple occupancies
	3	Public transportation
	4	Non-motorized
	5	Telecommute
polview	1	Extremely liberal
	2	Liberal
	3	Slightly liberal
	4	Moderate
	5	Slightly conservative
	6	Conservative
	7	Extremely conservative
polparty	0	No
	1	Yes
polcontrib	0	No
	1	Yes
vote	0	No
	1	Yes
card	1	American Express
	2	Visa
	3	Mastercard
	4	Discover
	5	Other
cardtype	1	None
	2	Gold
	3	Platinum
	4	Other
cardbenefit	1	None
	2	Cash back
	3	Airline miles
	4	Other
cardfee	0	No
	1	Yes
cardtenurecat	1	Less than 2
	2	2 to 5
	3	6 to 10
	4	11 to 15
	5	More than 15
owntv	0	No
	1	Yes

ownvcr	0	No
	1	Yes
owndvd	0	No
	1	Yes
owncd	0	No
	1	Yes
ownpda	0	No
	1	Yes
ownpc	0	No
	1	Yes
ownipod	0	No
	1	Yes
owngame	0	No
	1	Yes
ownfax	0	No
	1	Yes
news	0	No
	1	Yes

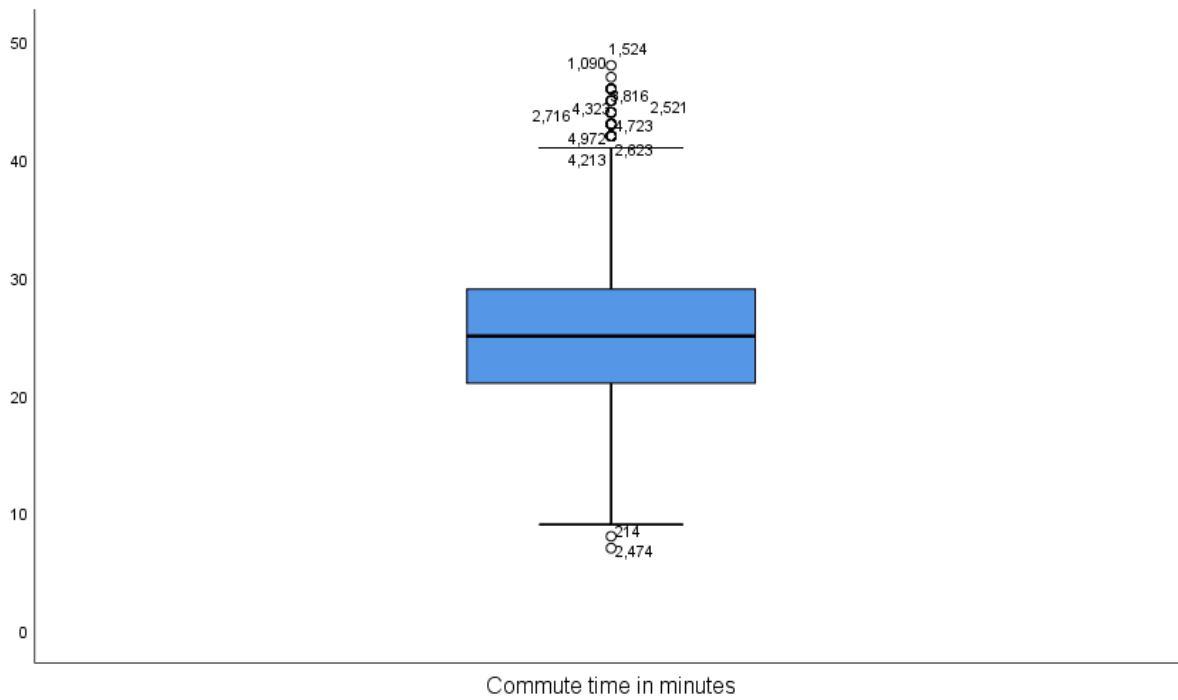
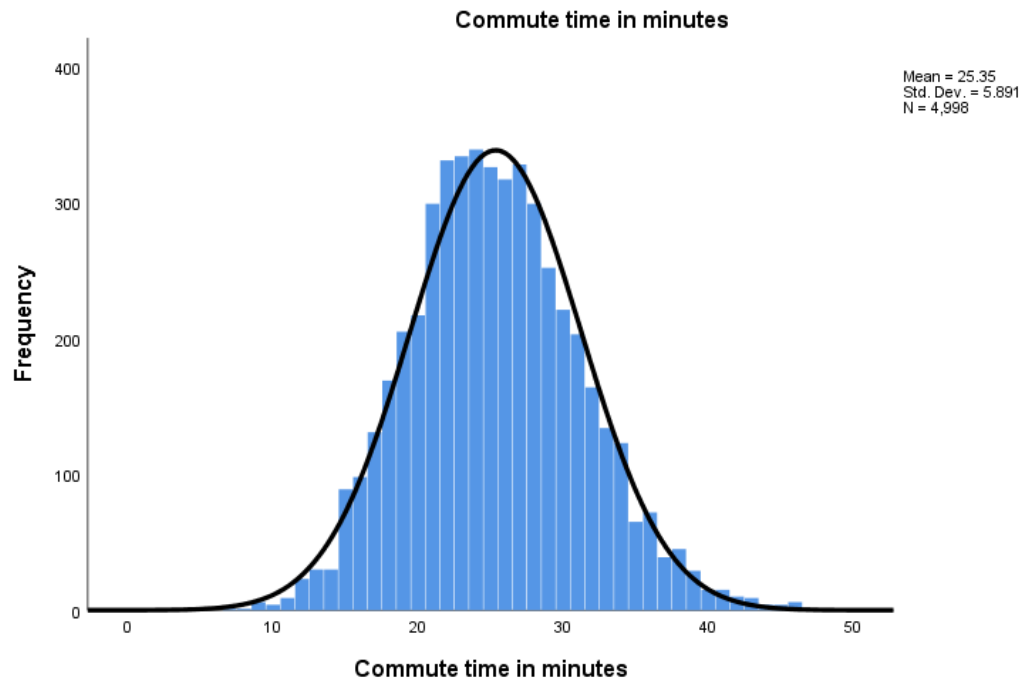
Annexure 3: Histograms and Boxplots

The following graphs provide more information about the distribution of the important continuous variables. This information should be used for making decisions about the prediction model.

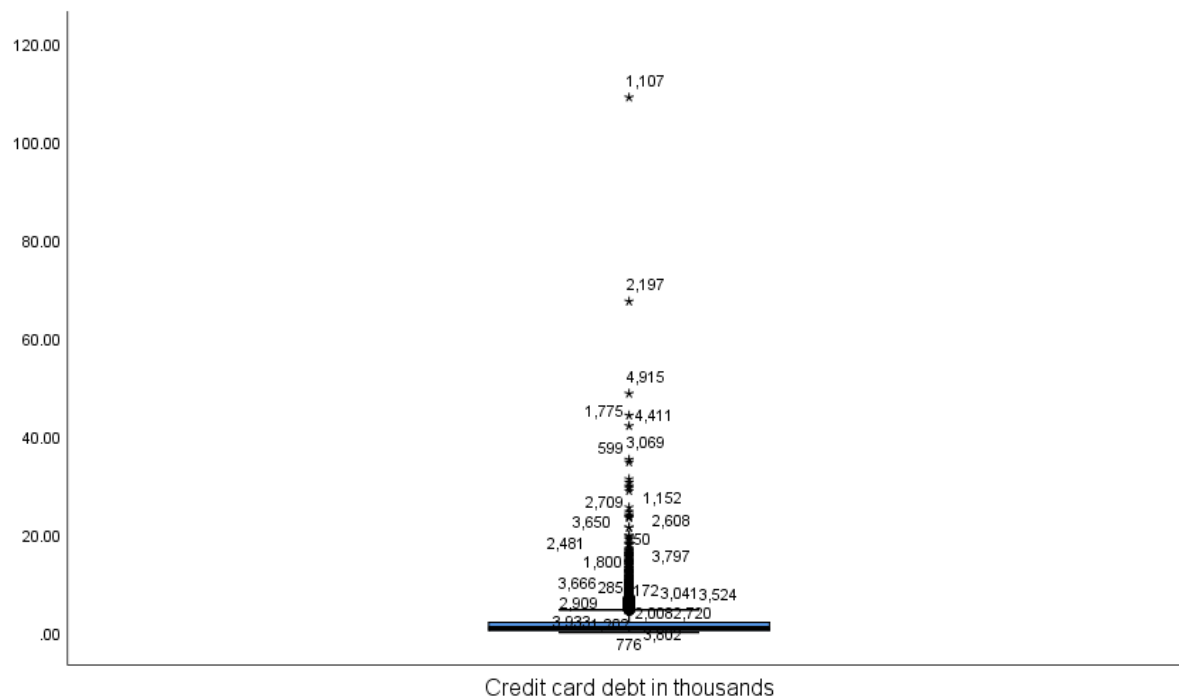
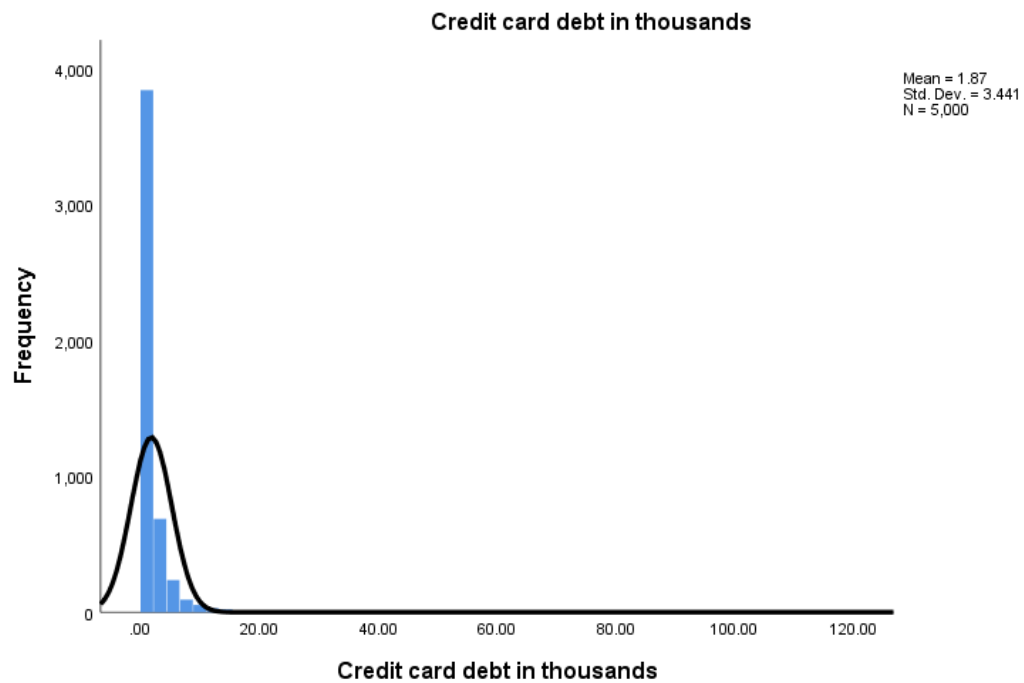
Age in years



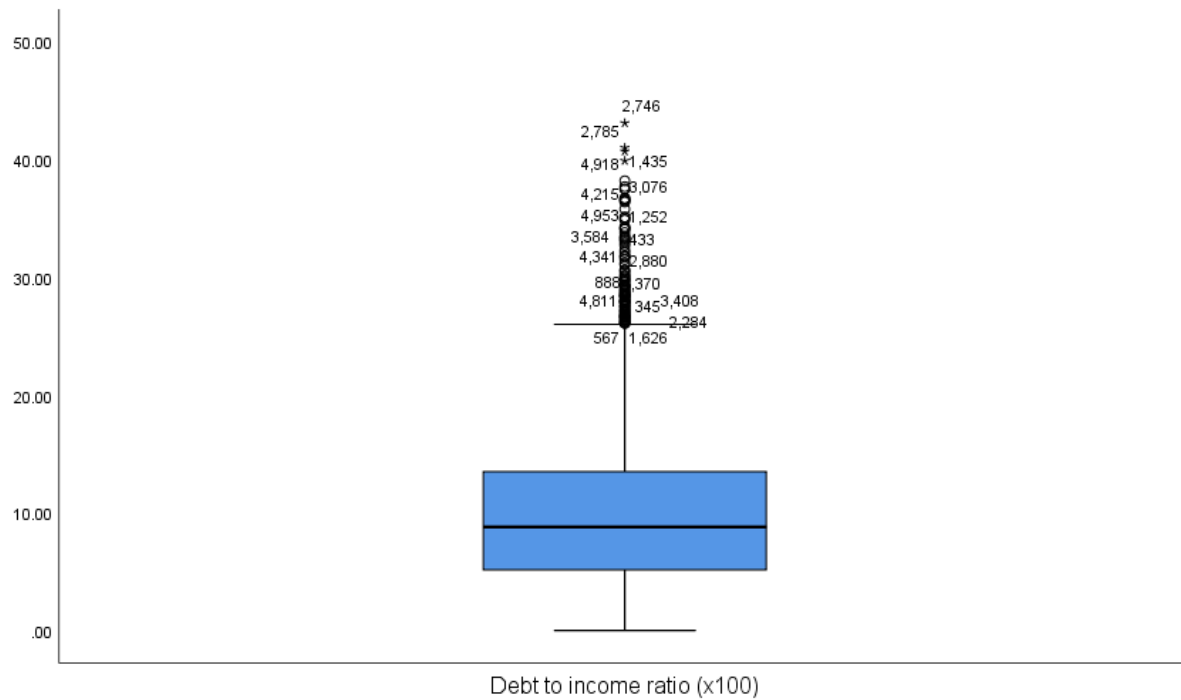
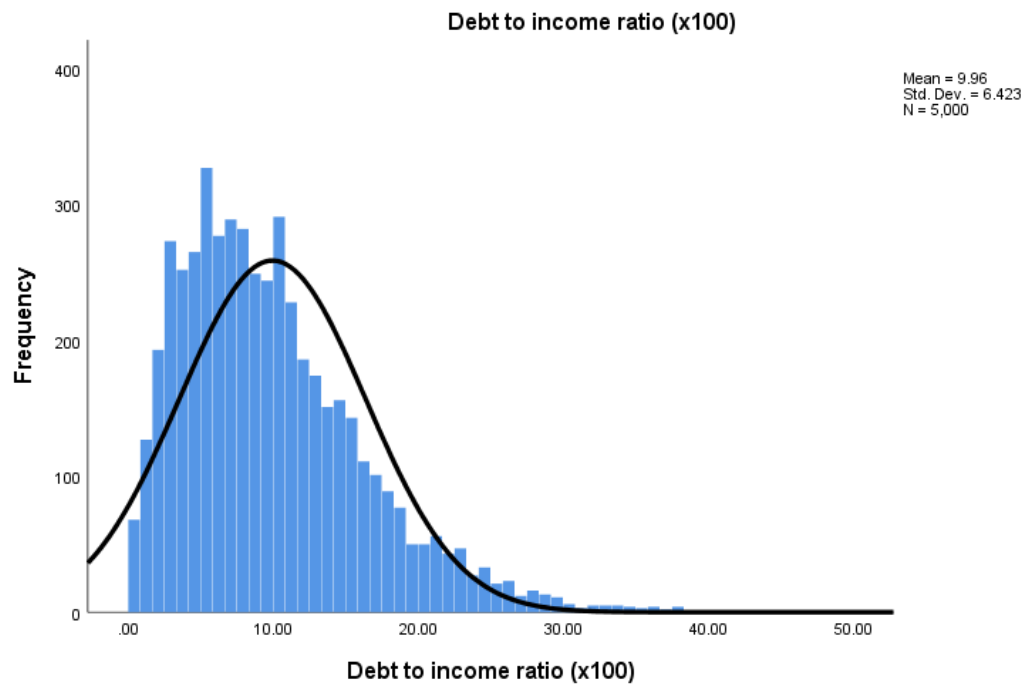
Commute time in minutes



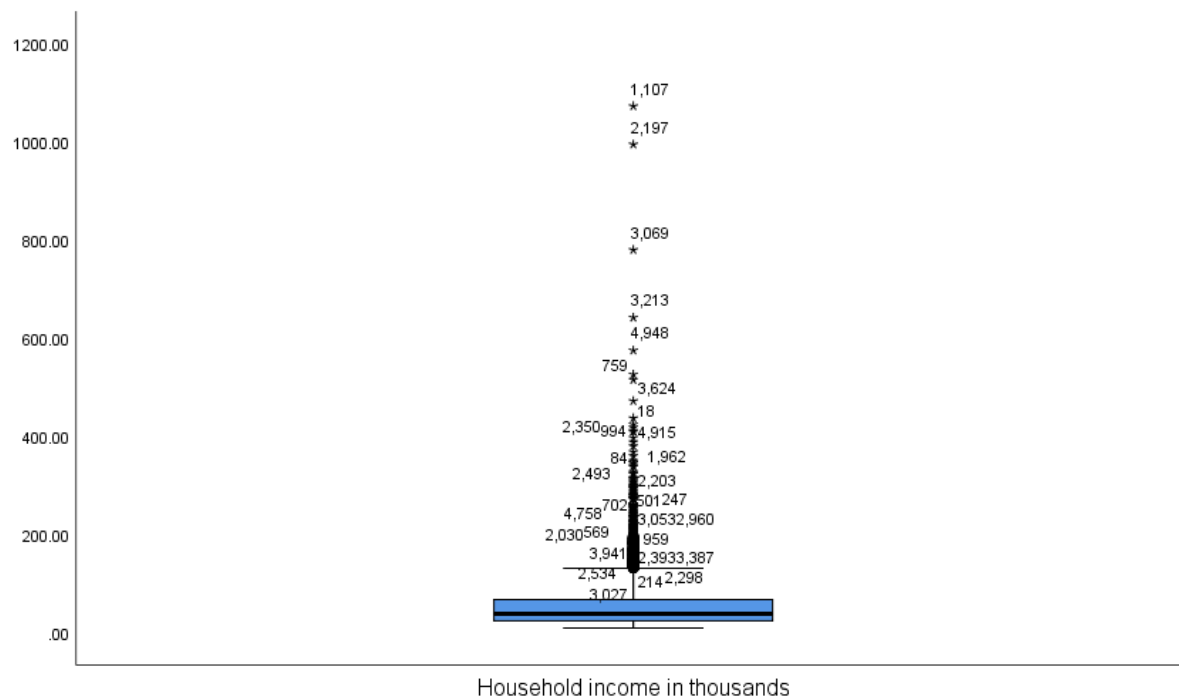
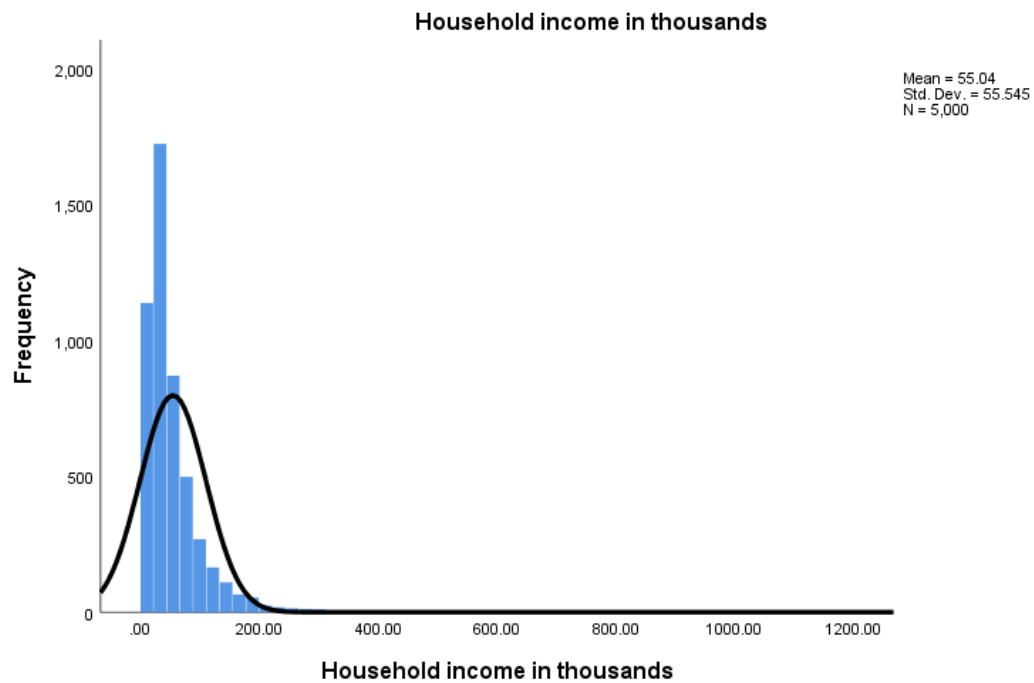
Credit card debt in thousands



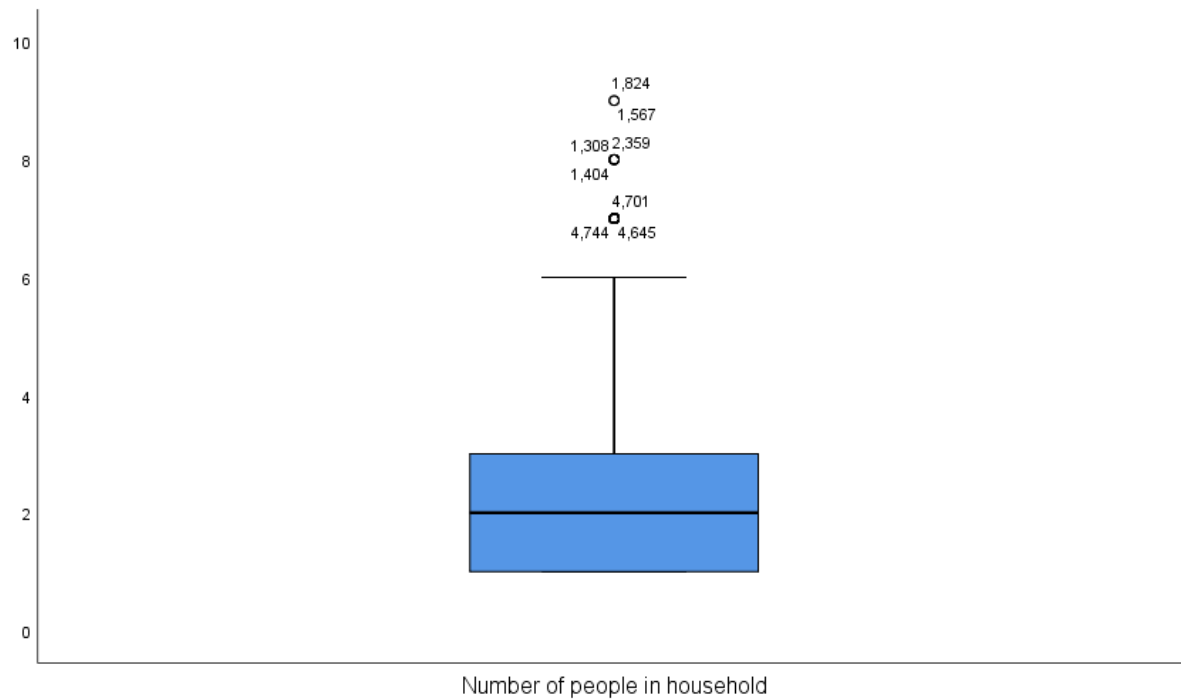
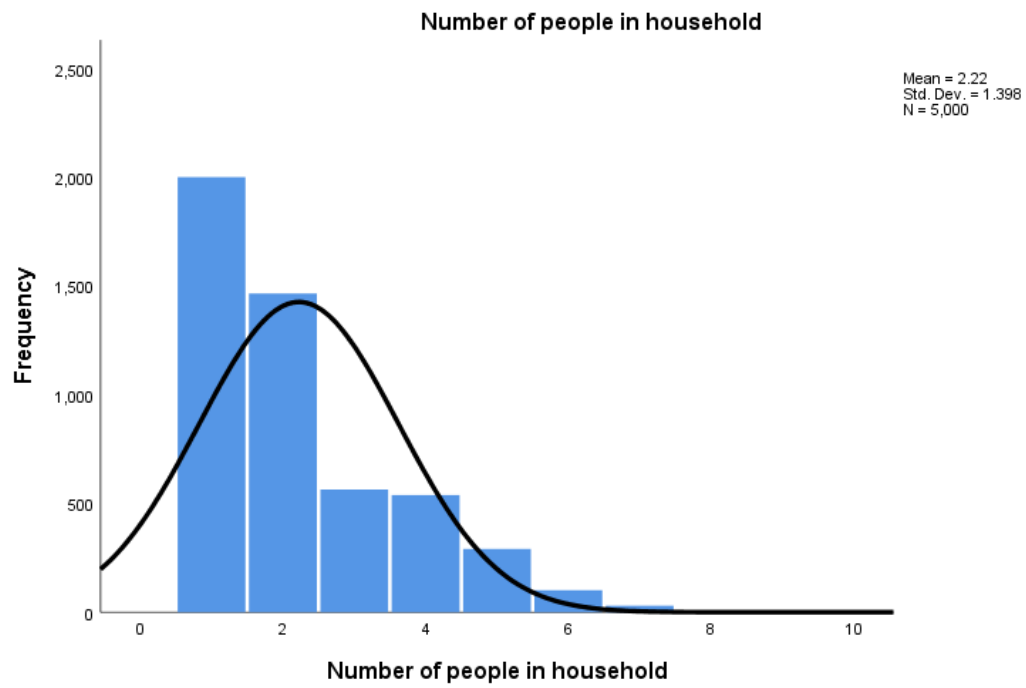
Debt to income ratio



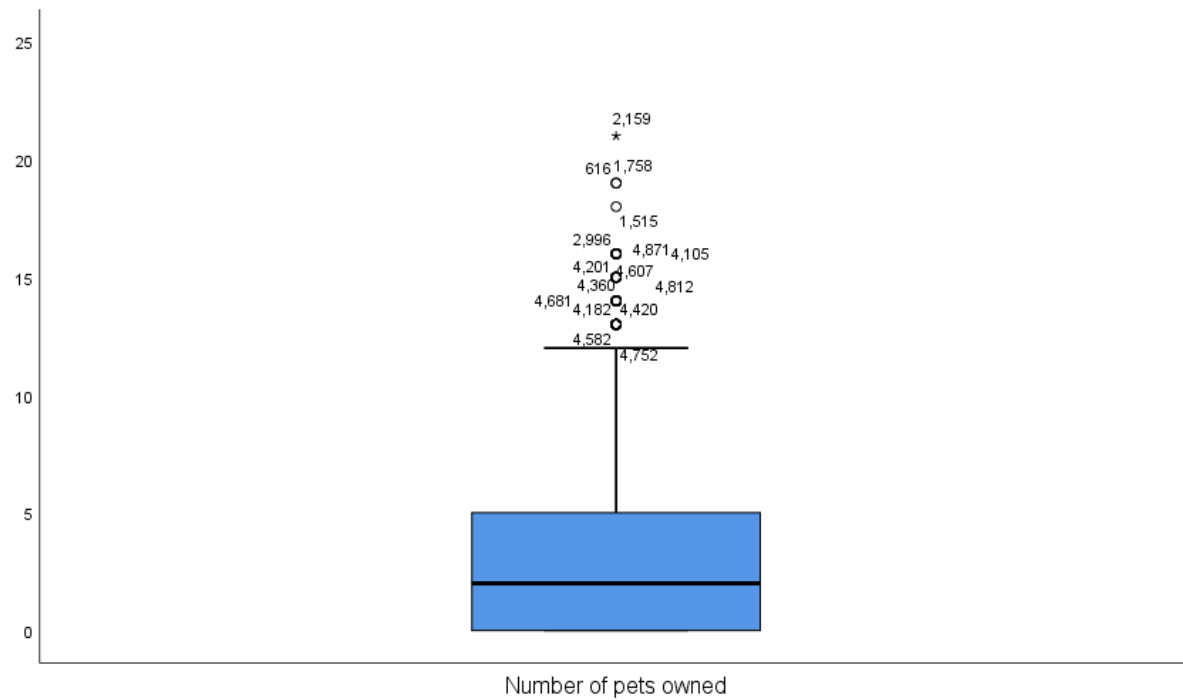
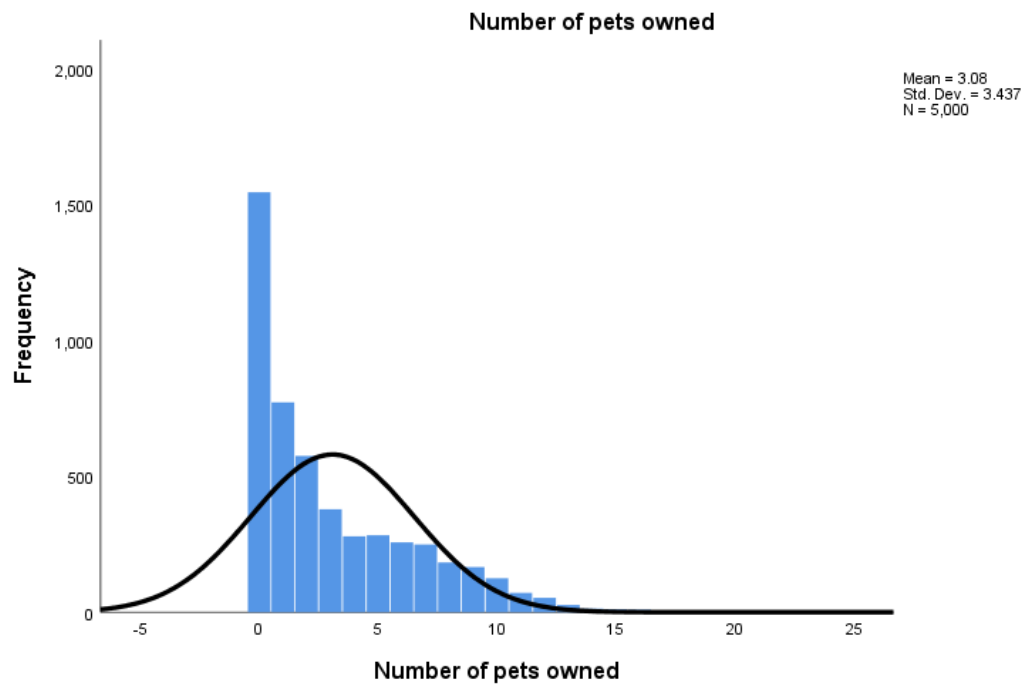
Household income in thousands



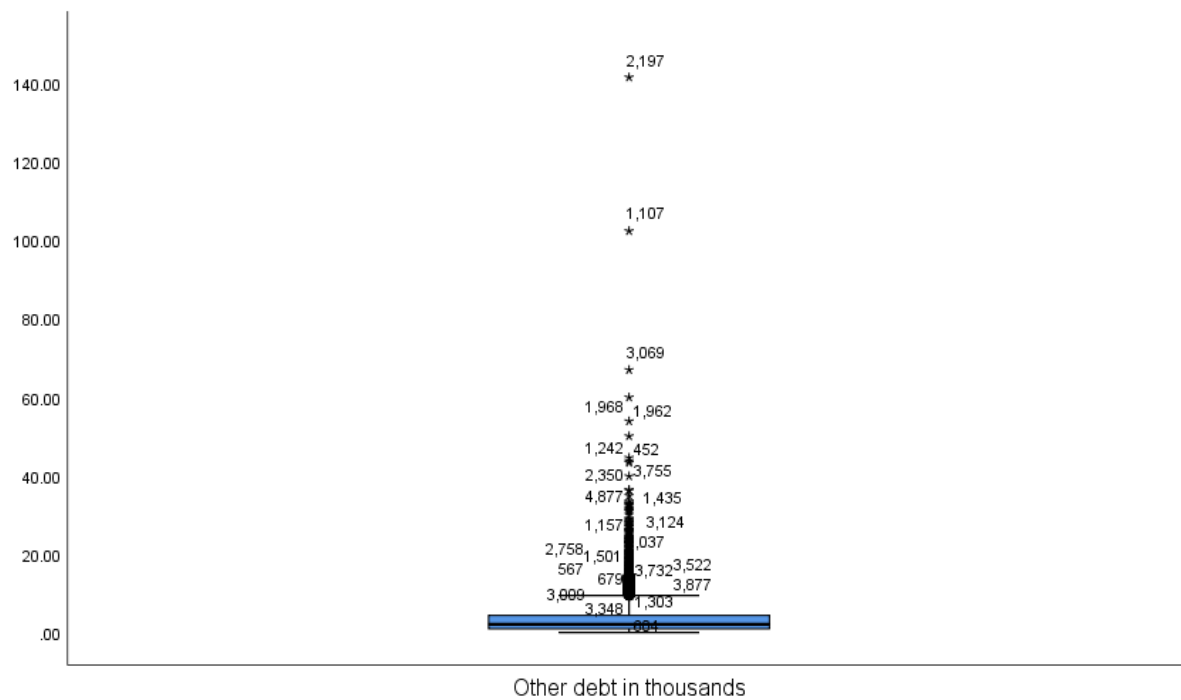
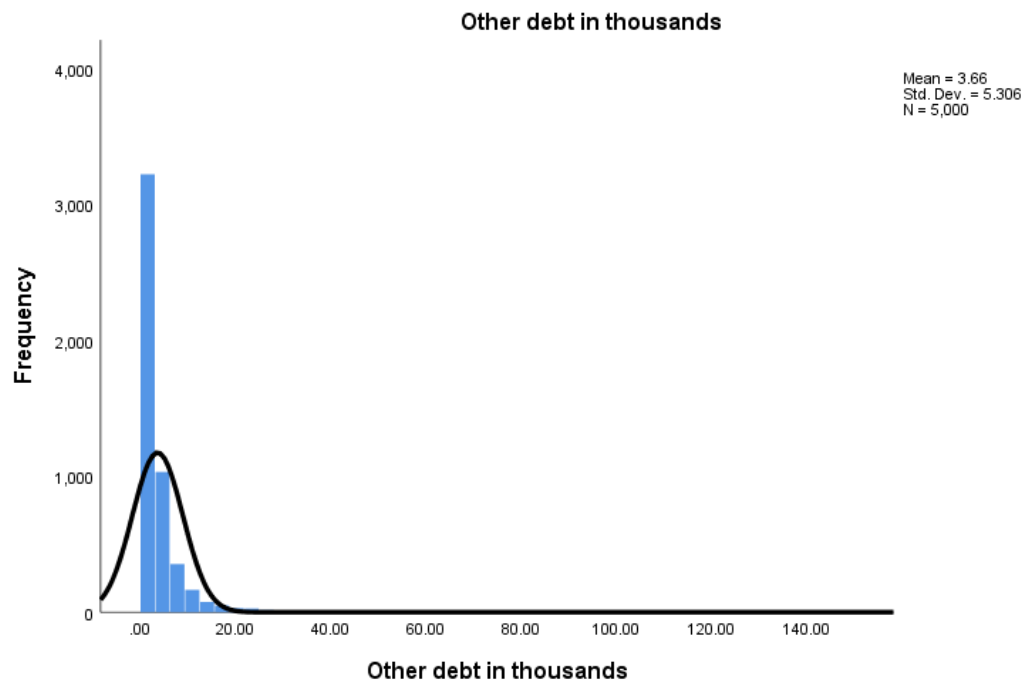
Number of people in household



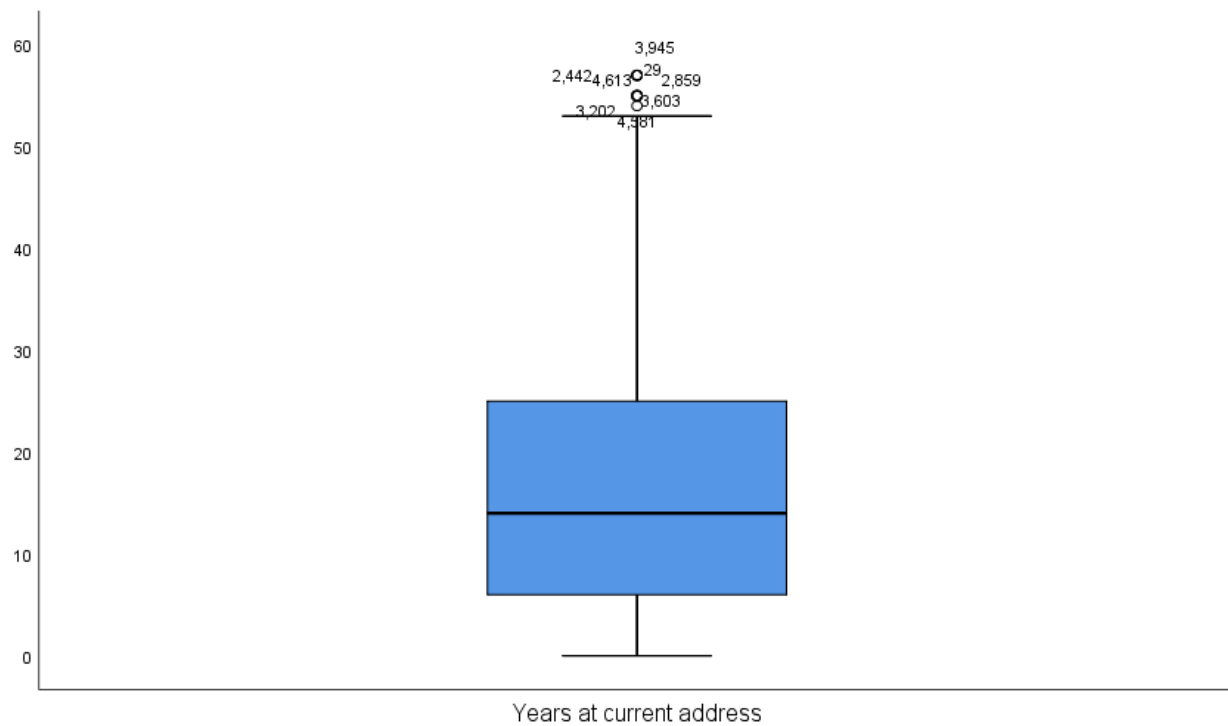
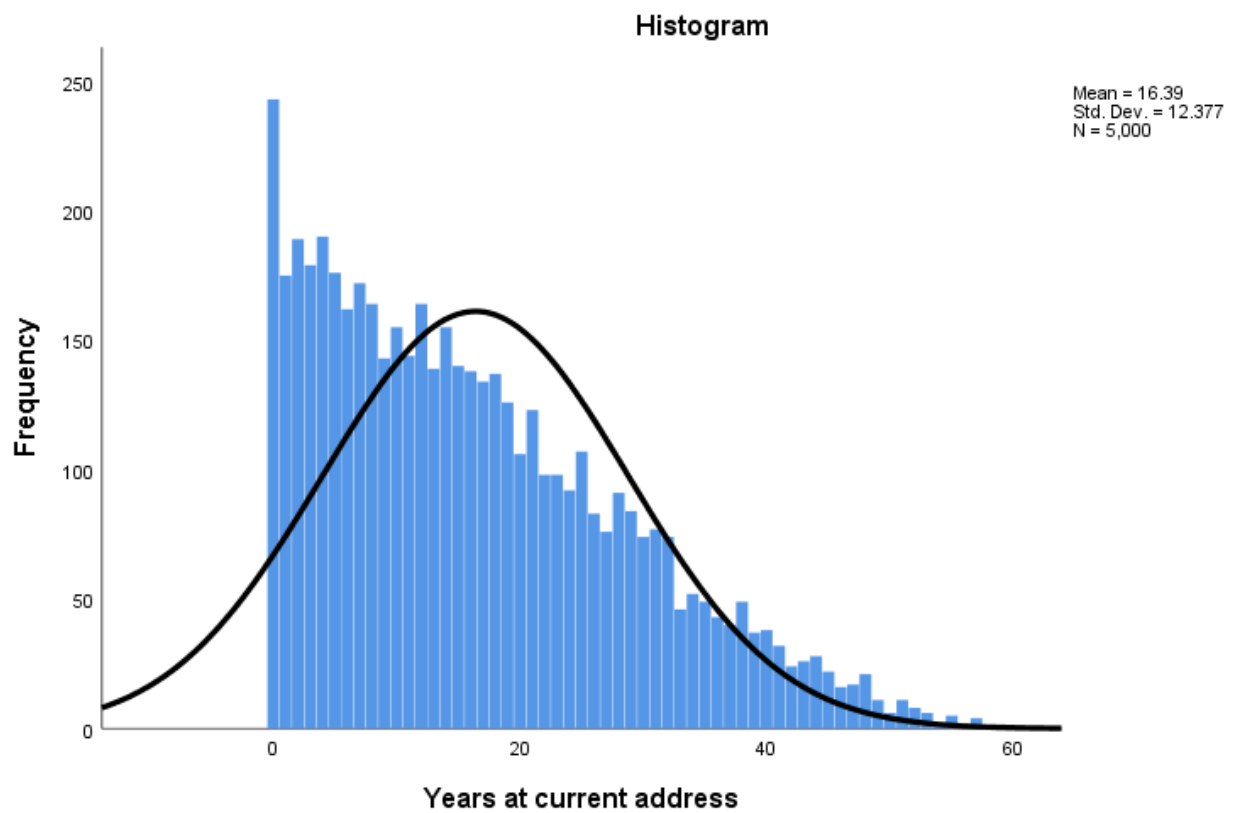
Number of pets owned



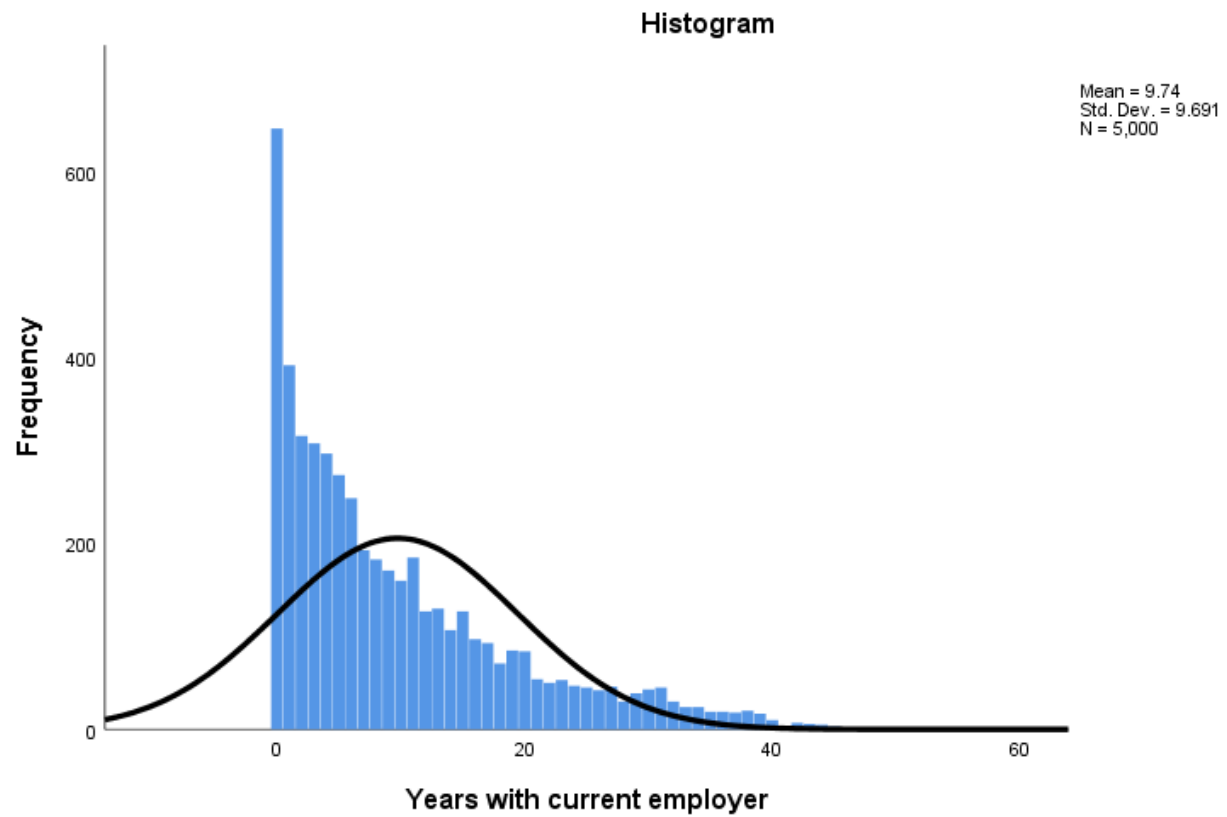
Other debt in thousands



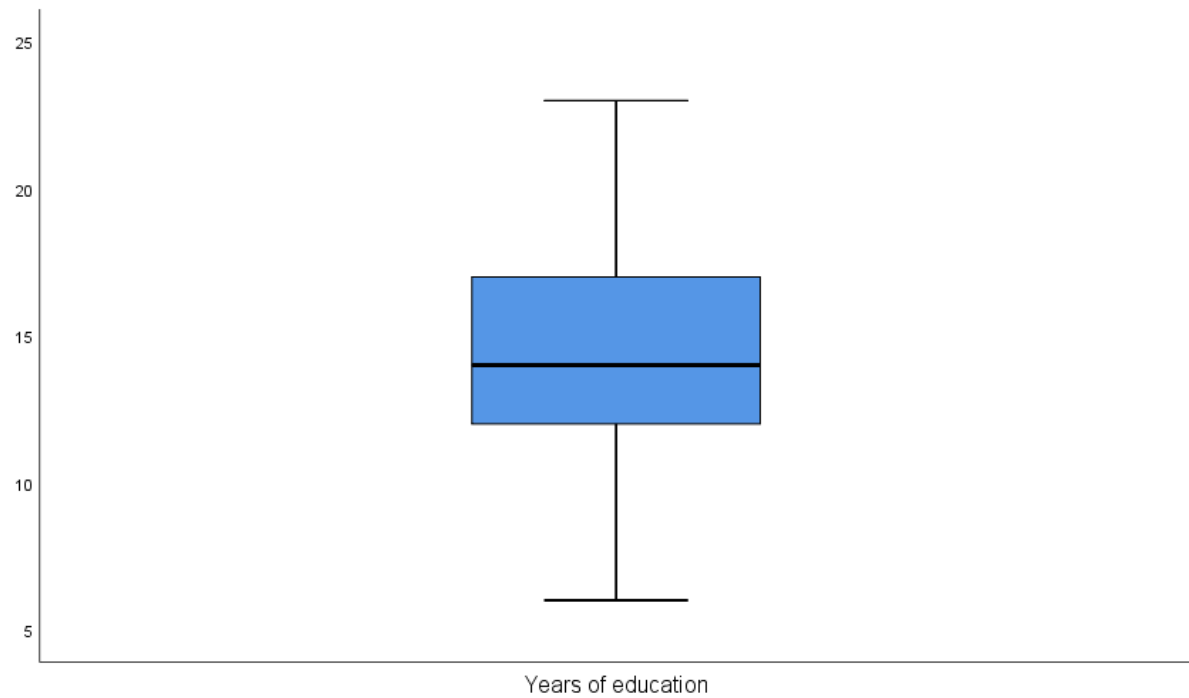
Years at current address



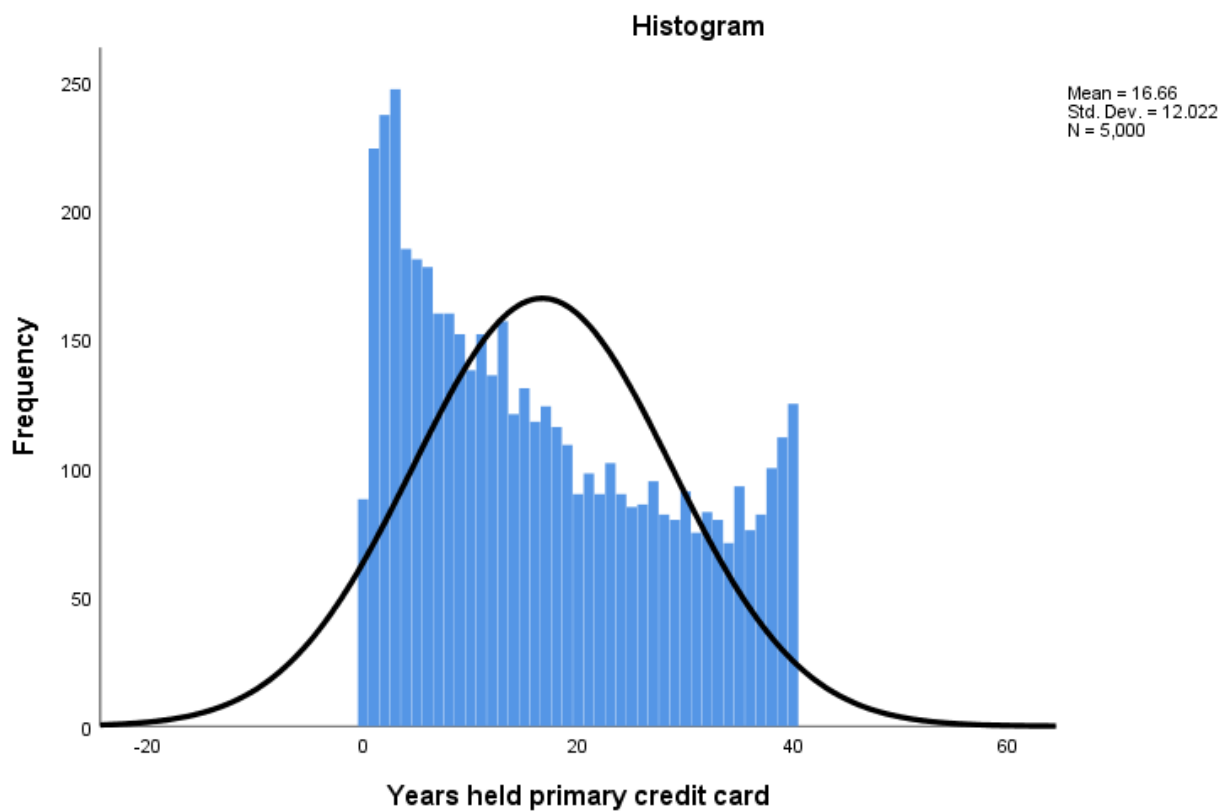
Years with current employer

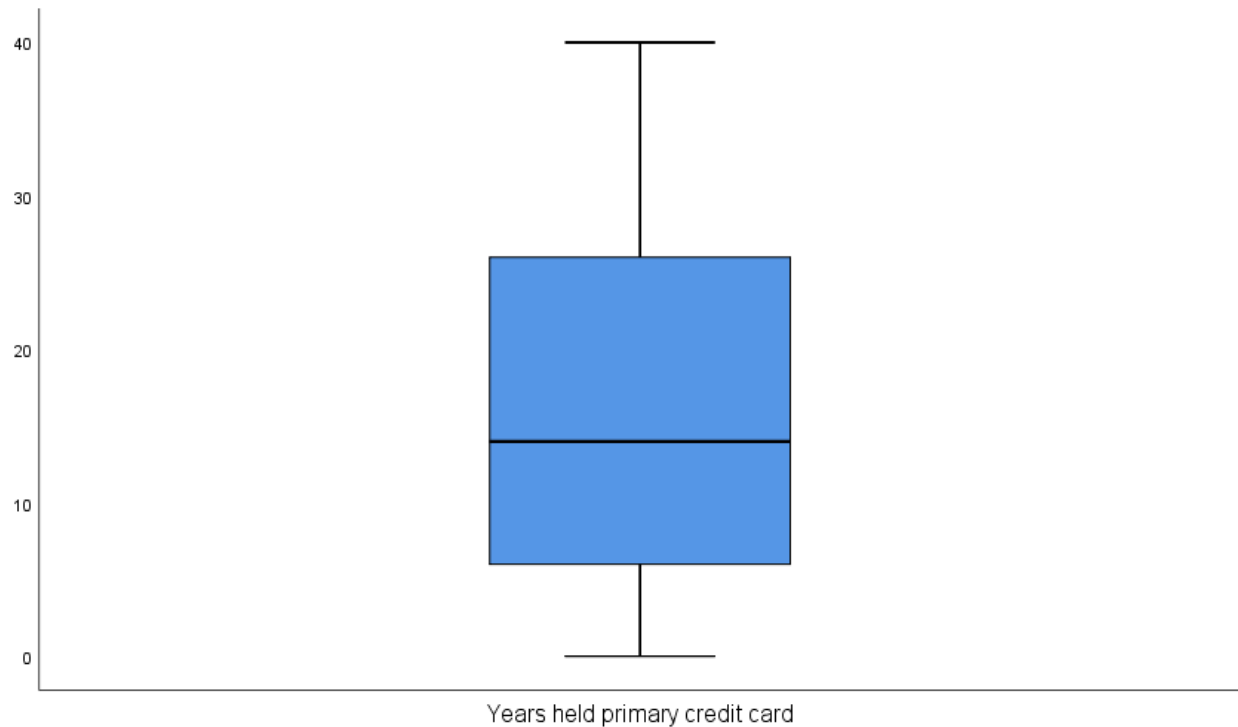






Years held the primary credit card





Annexure 4: Frequency distributions of categorical/ ordinal variables

cardfee Annual fee for primary credit card			
		Frequency	Percent
Valid	0 No	4050	81.0
	1 Yes	950	19.0
	Total	5000	100.0
cardbenefit Benefit program for primary credit card			
		Frequency	Percent
Valid	1 None	1249	25.0
	2 Cash back	1230	24.6
	3 Airline miles	1255	25.1
	4 Other	1266	25.3
	Total	5000	100.0
hometype Building type			
		Frequency	Percent
Valid	1 Single-family	2246	44.9
	2 Multiple-Family	1554	31.1
	3 Condominium/Townhouse	908	18.2
	4 Mobile Home	292	5.8

	Total	5000	100.0
commutecat Commute category			
		Frequency	Percent
Valid	1 Single occupancy	2917	58.3
	2 Multiple occupancies	296	5.9
	3 Public transportation	970	19.4
	4 Non-motorized	669	13.4
	5 Telecommute	148	3.0
	Total	5000	100.0
cardtype Designation of primary credit card			
		Frequency	Percent
Valid	1 None	1234	24.7
	2 Gold	1248	25.0
	3 Platinum	1255	25.1
	4 Other	1263	25.3
	Total	5000	100.0
gender Gender			
		Frequency	Percent
Valid	0 Male	2449	49.0
	1 Female	2551	51.0
	Total	5000	100.0
region Geographic indicator			
		Frequency	Percent
Valid	1 Zone 1	1019	20.4
	2 Zone 2	1005	20.1
	3 Zone 3	981	19.6
	4 Zone 4	943	18.9
	5 Zone 5	1052	21.0
	Total	5000	100.0
homeown Home ownership			
		Frequency	Percent
Valid	0 Rent	1846	36.9
	1 Own	3154	63.1
	Total	5000	100.0
jobcat Job category			

		Frequency	Percent
Valid	1 Managerial and Professional	1379	27.6
	2 Sales and Office	1630	32.6
	3 Service	628	12.6
	4 Agricultural and Natural Resources	218	4.4
	5 Precision Production, Craft, Repair	446	8.9
	6 Operation, Fabrication, General Labor	699	14.0
	Total	5000	100.0
marital Marital status			
		Frequency	Percent
Valid	0 Unmarried	2559	51.2
	1 Married	2441	48.8
	Total	5000	100.0
news Newspaper subscription			
		Frequency	Percent
Valid	0 No	2633	52.7
	1 Yes	2367	47.3
	Total	5000	100.0
ownpc Owns computer			
		Frequency	Percent
Valid	0 No	1821	36.4
	1 Yes	3179	63.6
	Total	5000	100.0
owndvd Owns DVD player			
		Frequency	Percent
Valid	0 No	426	8.5
	1 Yes	4574	91.5
	Total	5000	100.0
ownfax Owns fax machine			
		Frequency	Percent
Valid	0 No	4087	81.7
	1 Yes	913	18.3
	Total	5000	100.0
owngame Owns gaming system			
		Frequency	Percent

Valid	0 No	2613	52.3
	1 Yes	2387	47.7
	Total	5000	100.0
ownpda Owns PDA			
		Frequency	Percent
Valid	0 No	3980	79.6
	1 Yes	1020	20.4
	Total	5000	100.0
ownipod Owns portable digital audio player			
		Frequency	Percent
Valid	0 No	2611	52.2
	1 Yes	2389	47.8
	Total	5000	100.0
owncd Owns stereo/CD player			
		Frequency	Percent
Valid	0 No	335	6.7
	1 Yes	4665	93.3
	Total	5000	100.0
owntv Owns TV			
		Frequency	Percent
Valid	0 No	85	1.7
	1 Yes	4915	98.3
	Total	5000	100.0
ownvcr Owns VCR			
		Frequency	Percent
Valid	0 No	421	8.4
	1 Yes	4579	91.6
	Total	5000	100.0
polcontrib Political contributions			
		Frequency	Percent
Valid	0 No	3850	77.0
	1 Yes	1150	23.0
	Total	5000	100.0
polparty Political party membership			

		Frequency	Percent
Valid	0 No	3105	62.1
	1 Yes	1895	37.9
	Total	5000	100.0
commute Primary commute transportation			
		Frequency	Percent
Valid	1 Car	2868	57.4
	2 Motorcycle	49	1.0
	3 Carpool	296	5.9
	4 Bus	629	12.6
	5 Train/Subway	298	6.0
	6 Other public transit	43	.9
	7 Bicycle	56	1.1
	8 Walk	584	11.7
	9 Other non-motorized transit	29	.6
	10 Telecommute	148	3.0
	Total	5000	100.0
card Primary credit card			
		Frequency	Percent
Valid	1 American Express	1000	20.0
	2 Visa	1250	25.0
	3 Mastercard	1202	24.0
	4 Discover	1333	26.7
	5 Other	215	4.3
	Total	5000	100.0
retire Retired			
		Frequency	Percent
Valid	0 No	4268	85.4
	1 Yes	732	14.6
	Total	5000	100.0
vote Voted in the last election			
		Frequency	Percent
Valid	0 No	2429	48.6
	1 Yes	2571	51.4
	Total	5000	100.0

agecat Age category			
		Frequency	Percent
Valid	2 18-24	623	12.5
	3 25-34	885	17.7
	4 35-49	1245	24.9
	5 50-64	1202	24.0
	6 >65	1045	20.9
	Total	5000	100.0
inccat Income category in thousands			
		Frequency	Percent
Valid	1 Under \$25	1330	26.6
	2 \$25 - \$49	1793	35.9
	3 \$50 - \$74	819	16.4
	4 \$75 - \$124	668	13.4
	5 \$125+	390	7.8
	Total	5000	100.0
jobsat Job satisfaction			
		Frequency	Percent
Valid	1 Highly dissatisfied	967	19.3
	2 Somewhat dissatisfied	1041	20.8
	3 Neutral	1092	21.8
	4 Somewhat satisfied	1014	20.3
	5 Highly satisfied	886	17.7
	Total	5000	100.0
edcat Level of education			
		Frequency	Percent
Valid	1 Did not complete high school	953	19.1
	2 High school degree	1571	31.4
	3 Some college	1002	20.0
	4 College degree	1113	22.3
	5 Post-undergraduate degree	361	7.2
	Total	5000	100.0
cars Number of cars owned/leased			
		Frequency	Percent
Valid	0	492	9.8

	1	1113	22.3
	2	1620	32.4
	3	1076	21.5
	4	485	9.7
	5	150	3.0
	6	49	1.0
	7	14	.3
	8	1	.0
	Total	5000	100.0

polview Political outlook

		Frequency	Percent
Valid	1 Extremely liberal	163	3.3
	2 Liberal	610	12.2
	3 Slightly liberal	648	13.0
	4 Moderate	1763	35.3
	5 Slightly conservative	888	17.8
	6 Conservative	842	16.8
	7 Extremely conservative	86	1.7
	Total	5000	100.0

townsize Size of hometown

		Frequency	Percent
Valid	1 > 250,000	1430	28.6
	2 50,000-249,999	1055	21.1
	3 10,000-49,999	896	17.9
	4 2,500-9,999	861	17.2
	5 < 2,500	756	15.1
	Total	4998	100.0
Missing	System	2	.0
Total		5000	100.0

addresscat Years at current address

		Frequency	Percent
Valid	1 Less than 3	607	12.1
	2 4 to 7	879	17.6
	3 8 to 15	1204	24.1
	4 16 to 25	1159	23.2
	5 More than 25	1151	23.0

	Total	5000	100.0
cardtenurecat Years held the primary credit card			
		Frequency	Percent
Valid	1 Less than 2	312	6.2
	2 2 to 5	850	17.0
	3 6 to 10	788	15.8
	4 11 to 15	697	13.9
	5 More than 15	2353	47.1
	Total	5000	100.0
empcat Years with current employer			
		Frequency	Percent
Valid	1 Less than 2	1039	20.8
	2 2 to 5	1195	23.9
	3 6 to 10	956	19.1
	4 11 to 15	676	13.5
	5 More than 15	1134	22.7
	Total	5000	100.0