MTH 324, Final Exam, At Home, Fall 2022 Name KEY

Instructions: This exam is in two parts: Part I is to be completed partly at home using the materials posted in the course for the at-home portion and you will answer questions about that work during the in-class portion of the exam; Part II is to be completed entirely in class. You may not use cell phones, and you may only access internet resources you are specifically directed to use.

At home, prepare for questions in Part I using R. Open the data file entitled **324exam2data.xlsx** posted in Blackboard. (Note: this file has multiple sheets of data. You may want to separate the data into separate files to upload to R, or look up how to access multiple sheets in R from a single upload.) Complete the calculations noted below. You will be asked for additional analysis and interpretation of this data in the in-class portion of the test. Print out the results of your analysis and code, and bring the pages with you to the exam. You will submit all this work along with the in-class exam.

1. Using the data on Sheet 1, conduct appropriate hypotheses of the paired data on ad preference. Is there sufficient reason to think one add is more preferred than the other? Test any assumptions of the hypothesis test you choose, and be prepared to explain your reasoning why you chose the test you did.

2. The data on Sheet 2 is complex with 12 variables (Person is not a variable).
   a. Explore the data. Make a graph of each variable. Use an appropriate graph for each variable type. For numerical variables, you'll need to be able to describe the shape of the distribution. Bar graphs or pie charts are appropriate for categorical data, or you can experiment with mosaic charts or others that you think display the individual variables well.

   b. Create summary statistics for the numerical variables.

   c. Create a Two-way table for Dwell Type and Neighborhood. Conduct a test of independence on this data.

   d. Create a comparative box plot of Age and Dwell Type. Conduct an ANOVA test on this data. If you reject the null, apply Tukey's method or pairwise comparisons as appropriate.

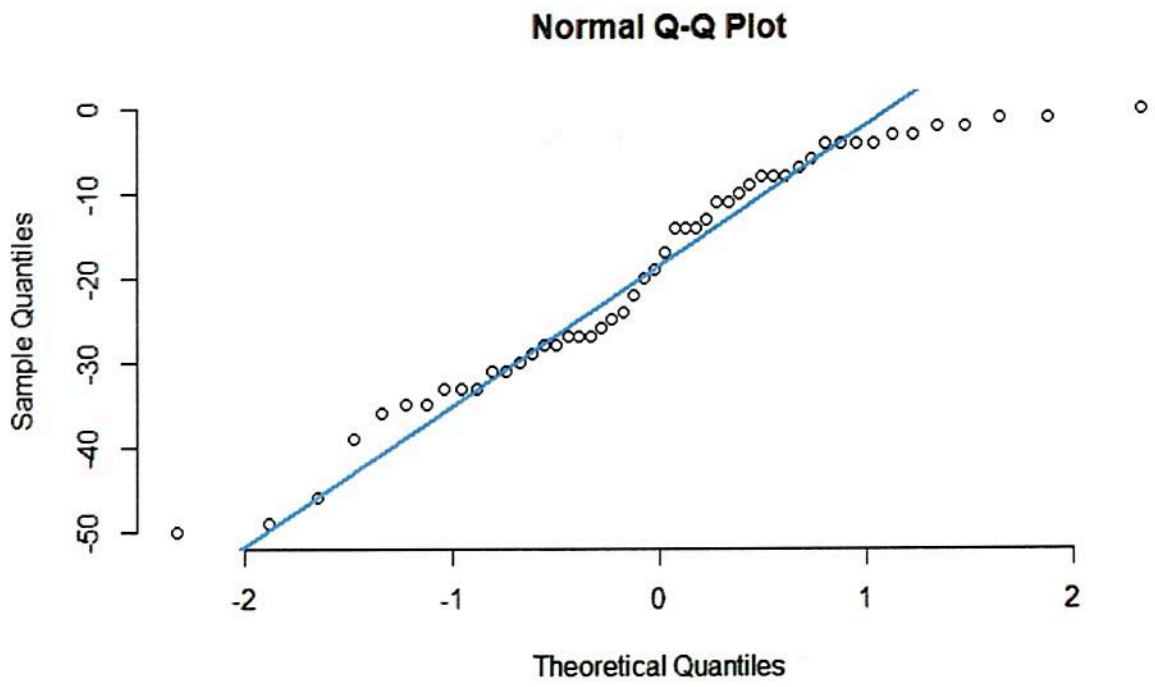
   e. Use bootstrapping to test whether there is good reason to think average credit card debt is less than 1500. Construct a confidence interval with this method, and using a traditional one-sample method. Compare the results.

   f. Construct a two-way table of Gender and Live Alone. Conduct a two-sample proportion test to see if the proportion of people living alone differs by gender.

3. The table below shows the results of a simulation of the sum of rolls of two pairs of dice using 250 trials.

	2	3	4	5	6	7	8	9	10	11	12
Count	11	10	22	21	35	41	31	29	25	19	6

   Dice rolls follow a triangle distribution, with $P(X = 2) = \frac{1}{36}$, which increases linearly to $P(X = 6) = P(X = 7) = \frac{6}{36}$, and then decreases linearly to $P(X = 12) = \frac{1}{36}$. Conduct a goodness-of-fit test to see if this data appears to follow this distribution.



Paired t-test

Paired t-test

data: x and y
t = -9.7819, df = 49, p-value = 4.164e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-23.16852 -15.27148
sample estimates:
mean of the differences
-19.22

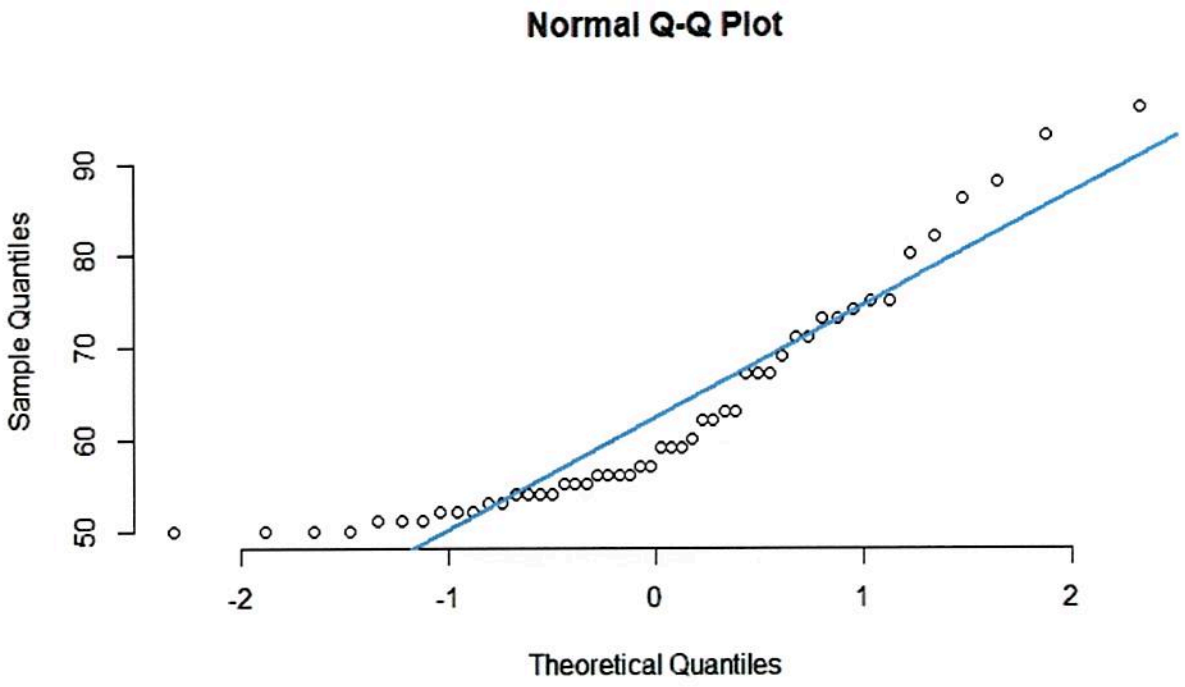
Wilcoxon test

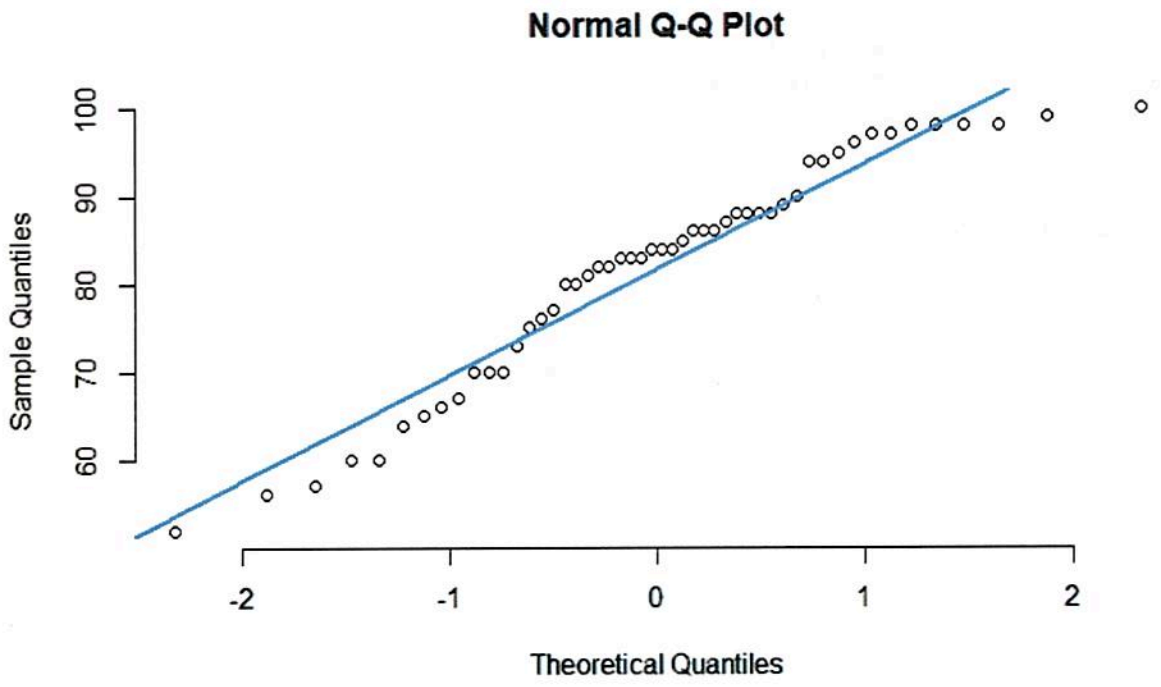
Wilcoxon signed rank test with continuity correction

data: x and y
V = 0, p-value = 1.136e-09
alternative hypothesis: true location shift is not equal to 0

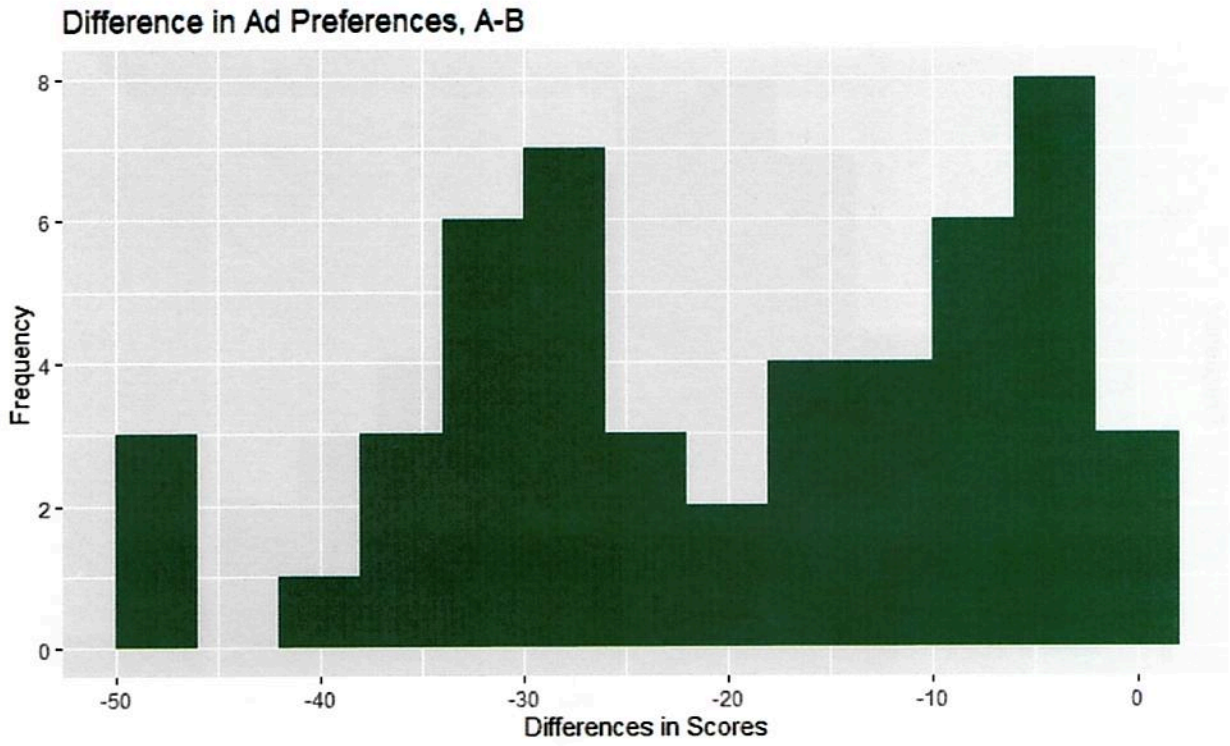
Final Exam Part 1 Work Solutions

1.

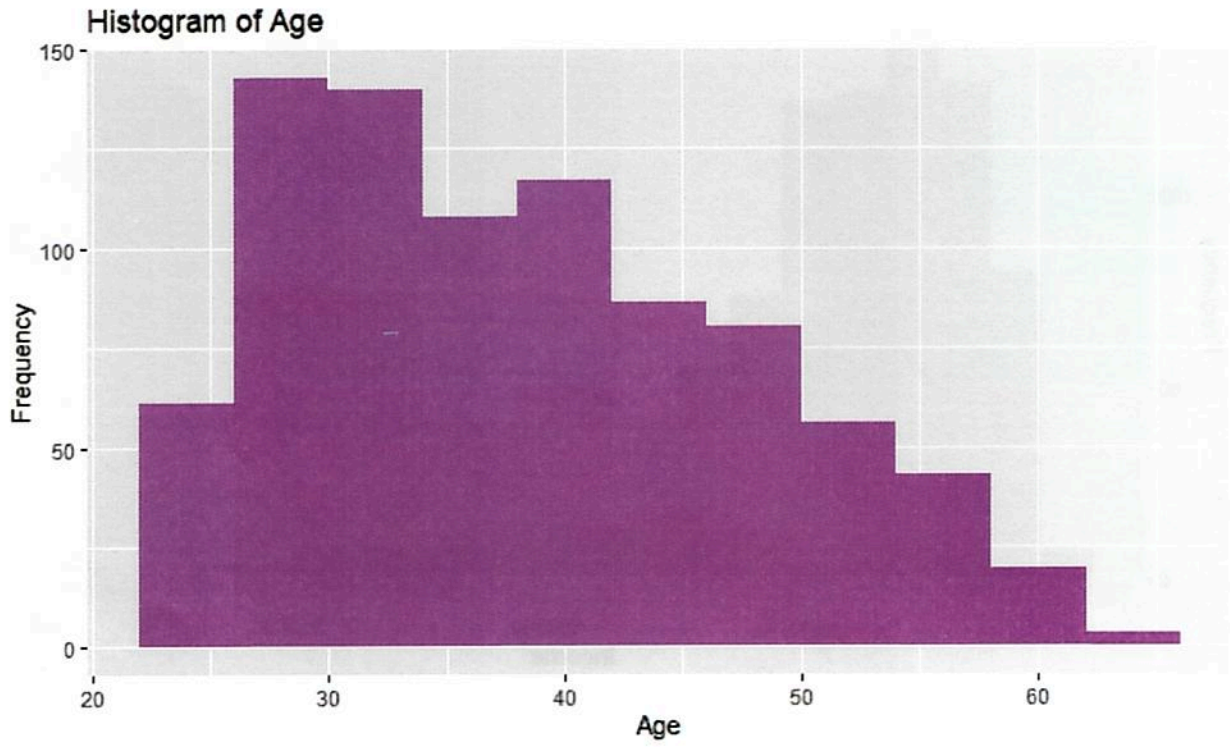


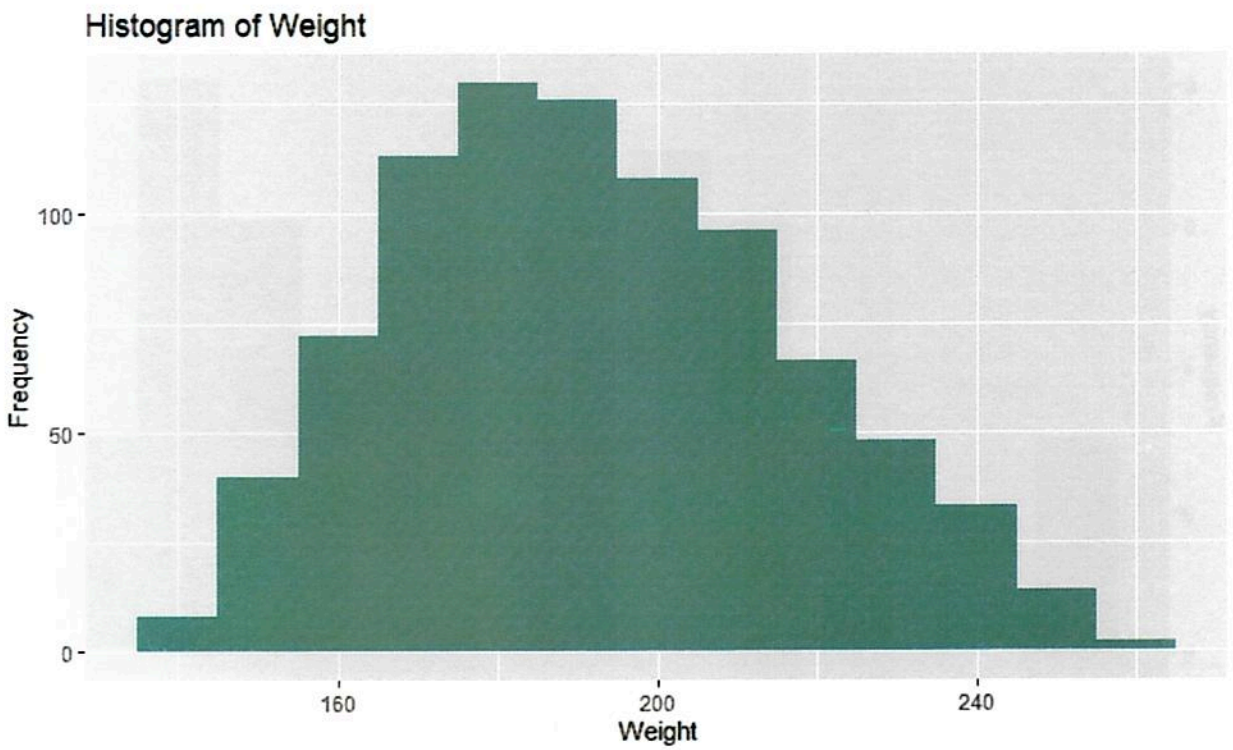


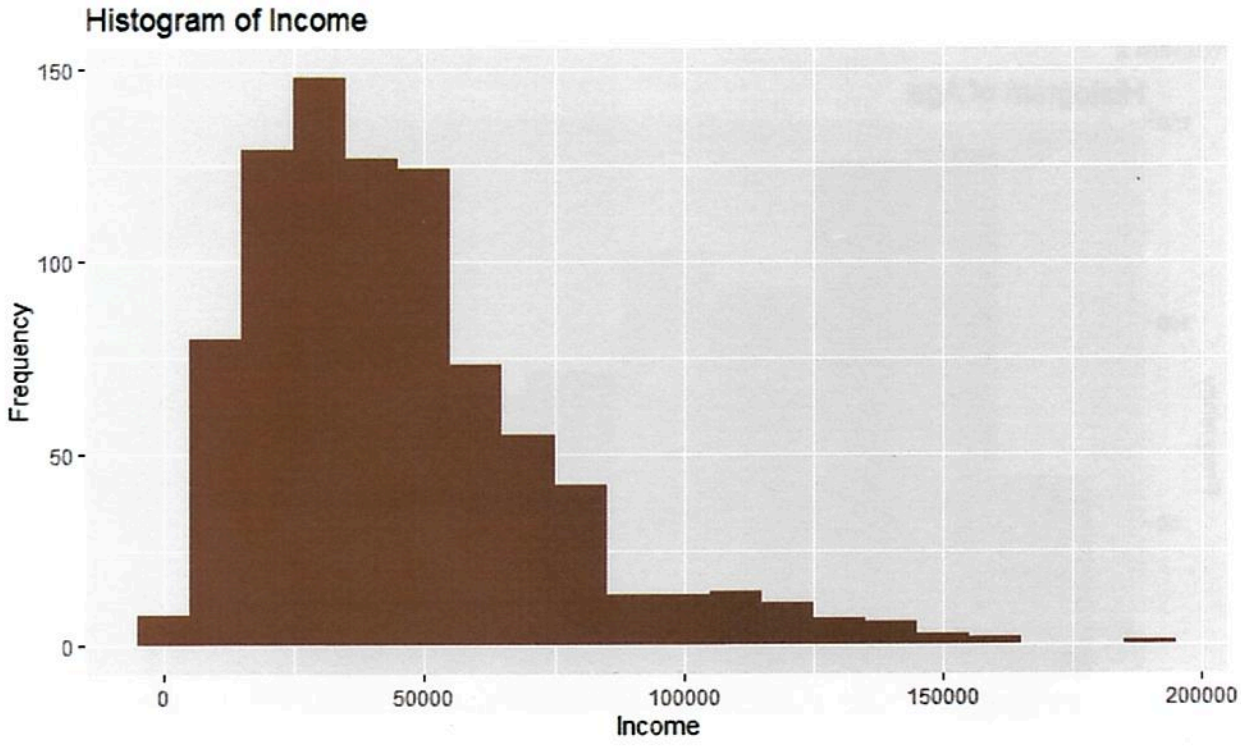
The first plot is Ad A, the second Ad B, the third is the differences.

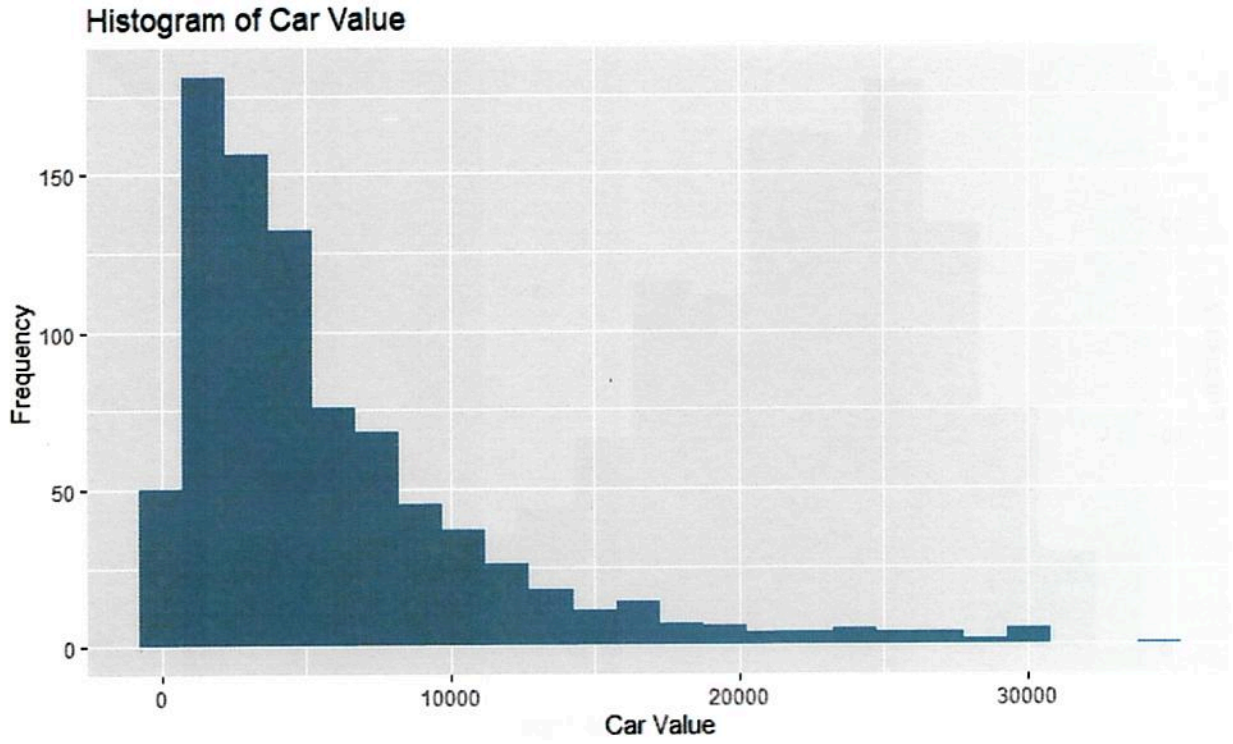

Difference in Ad Preferences, A-B
Frequency
8
6
4
2
0
-50
-40
-30
-20
-10
0
Differences in Scores

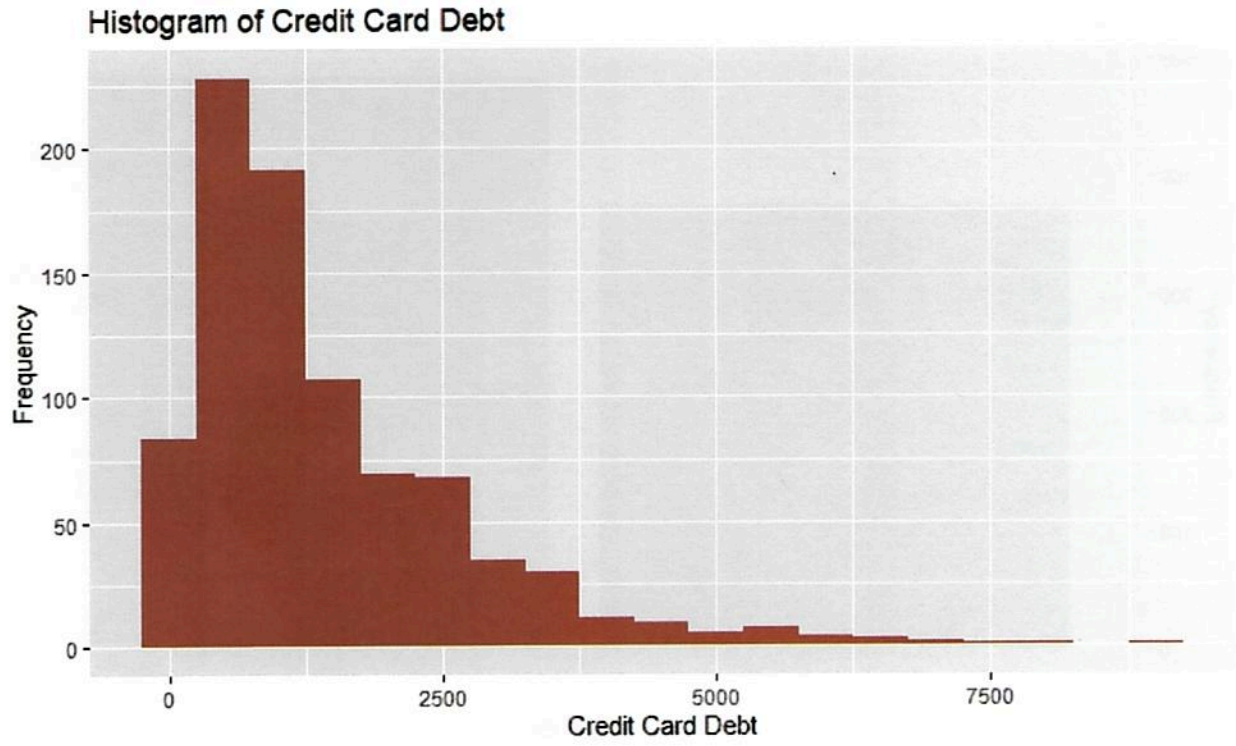
Problem 2

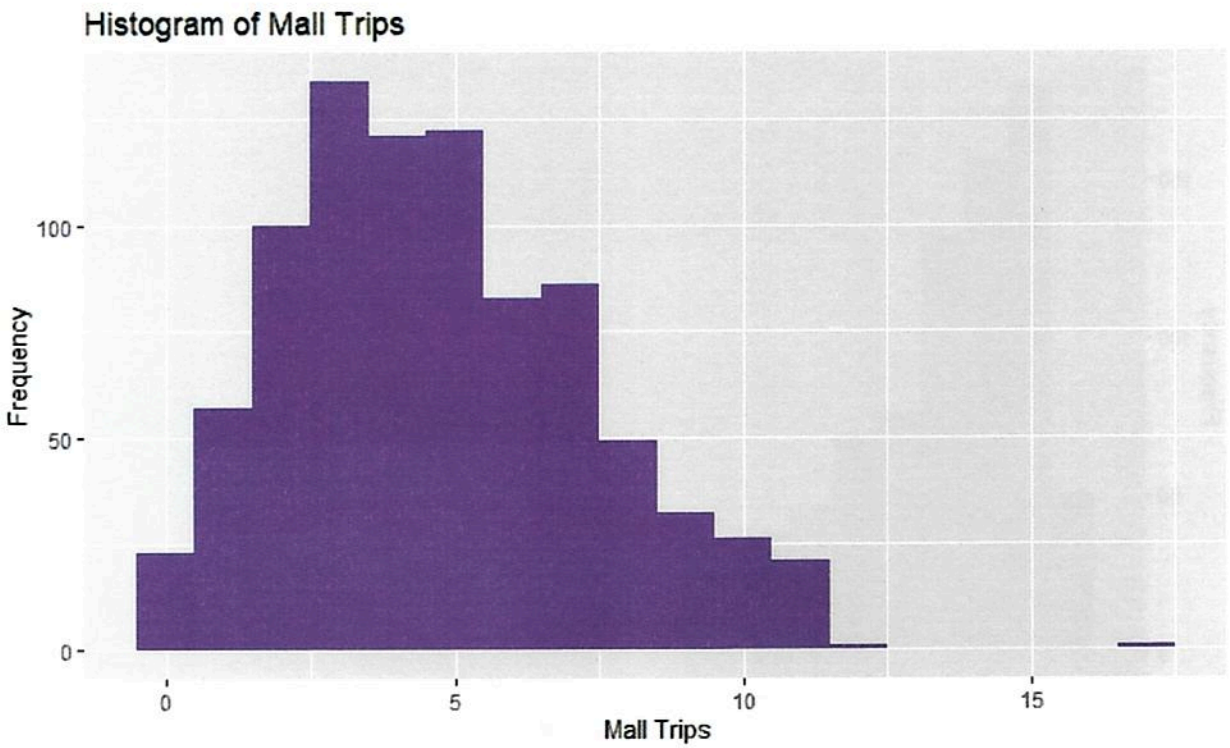

Histogram of Age
Frequency
150
100
50
0
20
30
40
50
60
Age

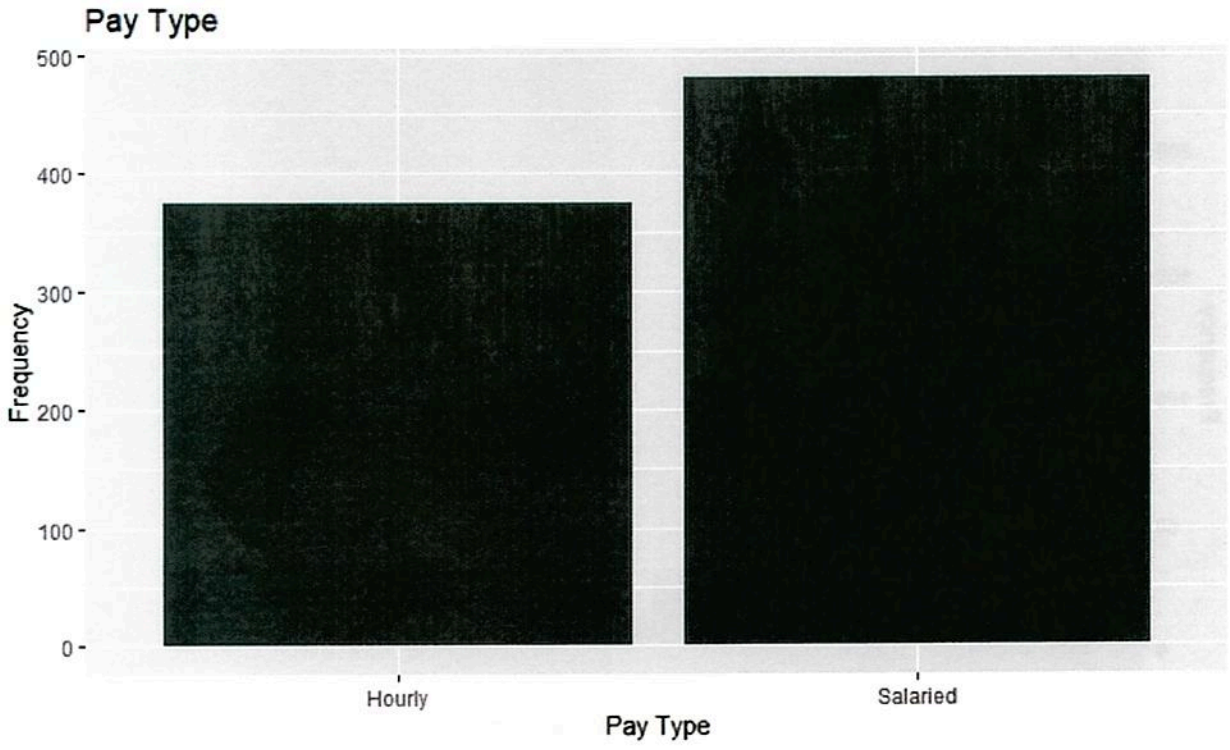

Histogram of Weight
Frequency
100
50
0
160
200
240
Weight

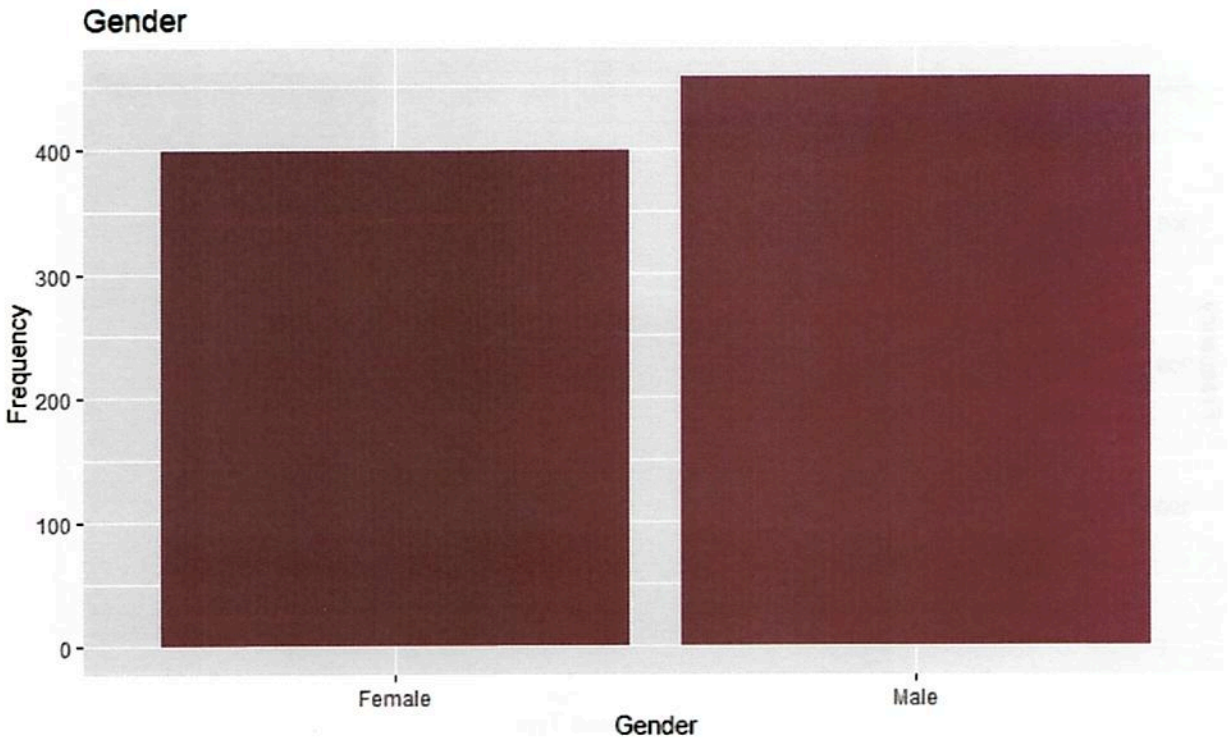

Histogram of Income
Frequency
150
100
50
0
0
50000
100000
150000
200000
Income

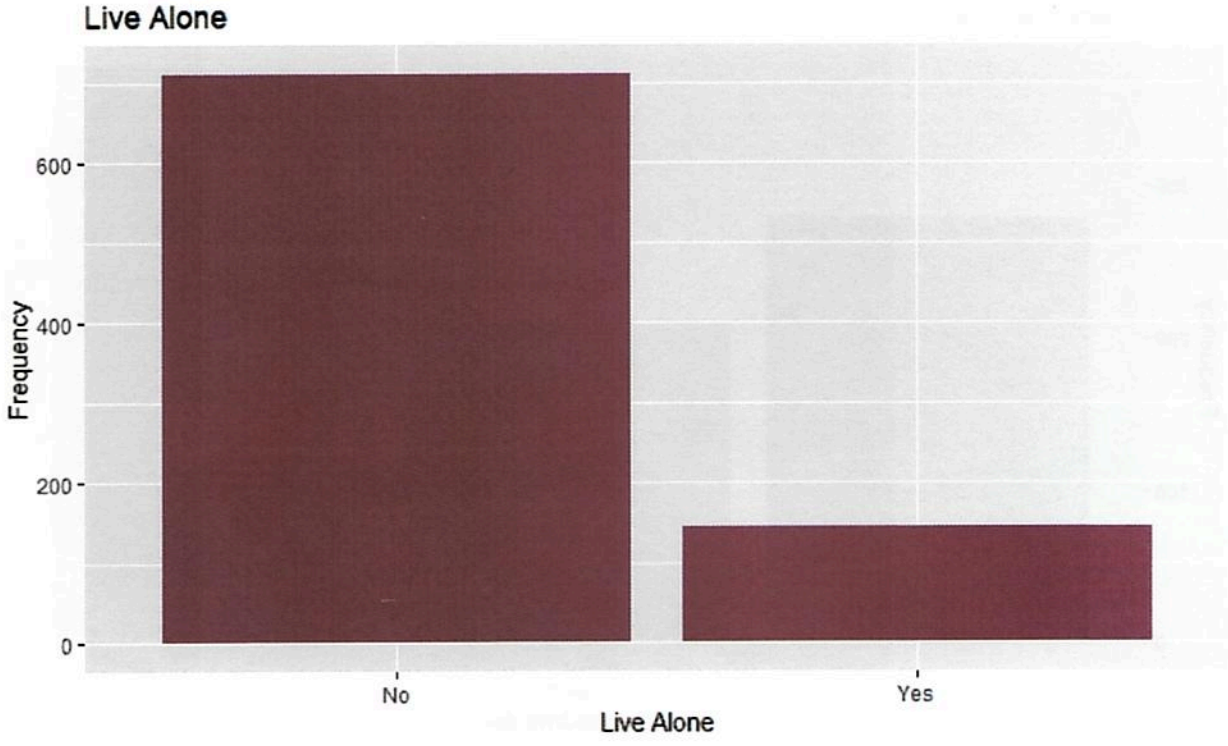

Histogram of Car Value
Frequency
150
100
50
0
0
10000
20000
30000
Car Value

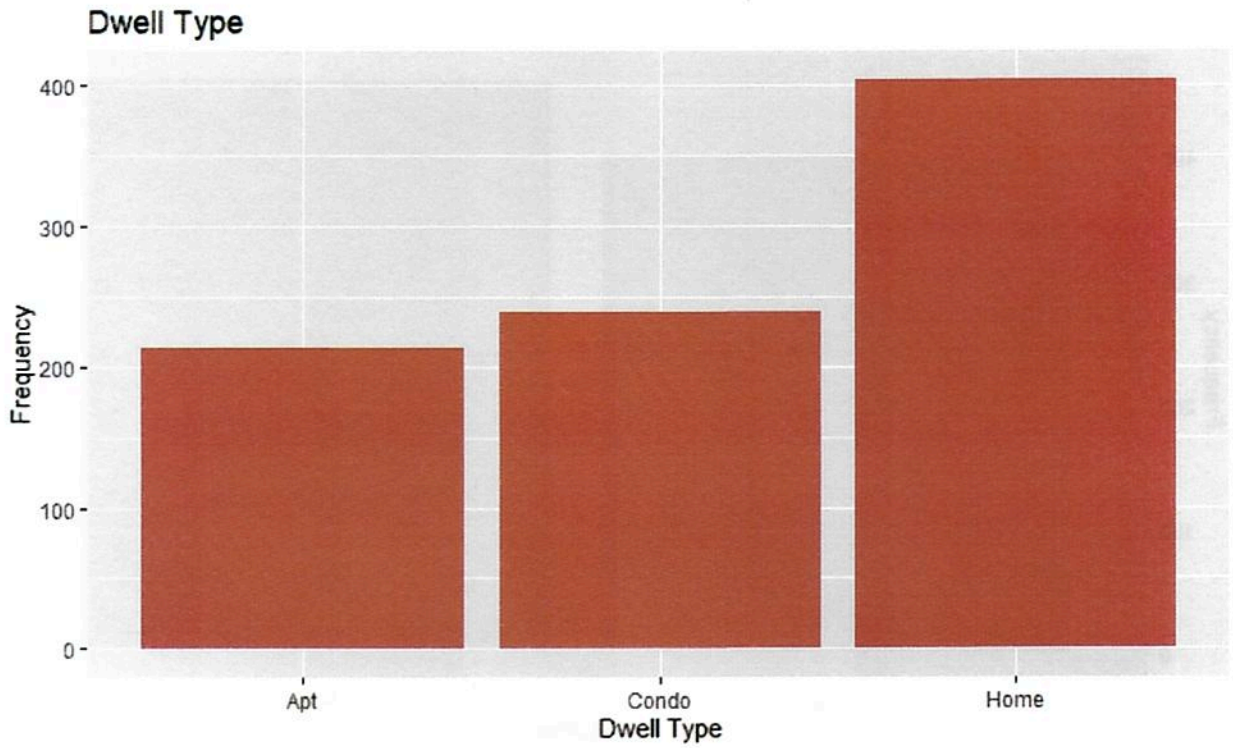

Histogram of Credit Card Debt
Frequency
200
150
100
50
0
0
2500
5000
7500
Credit Card Debt

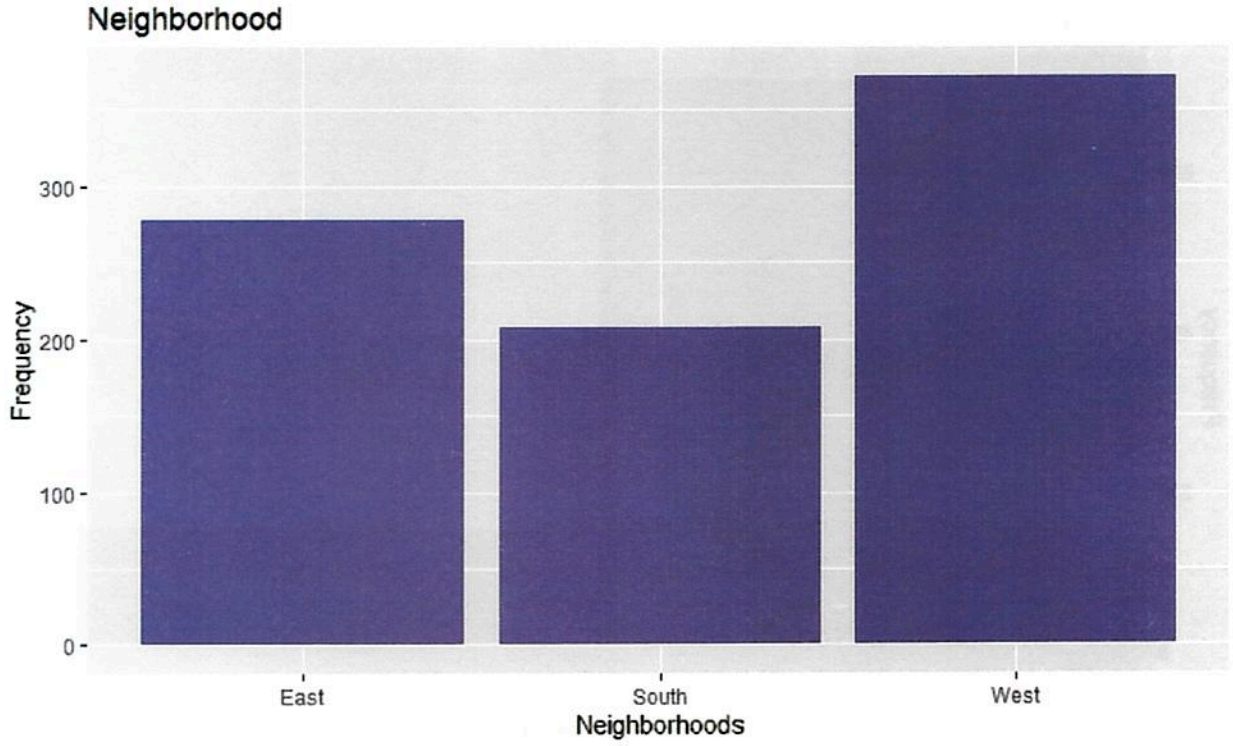

Histogram of Mall Trips
Frequency
100
50
0
0
5
10
15
Mall Trips

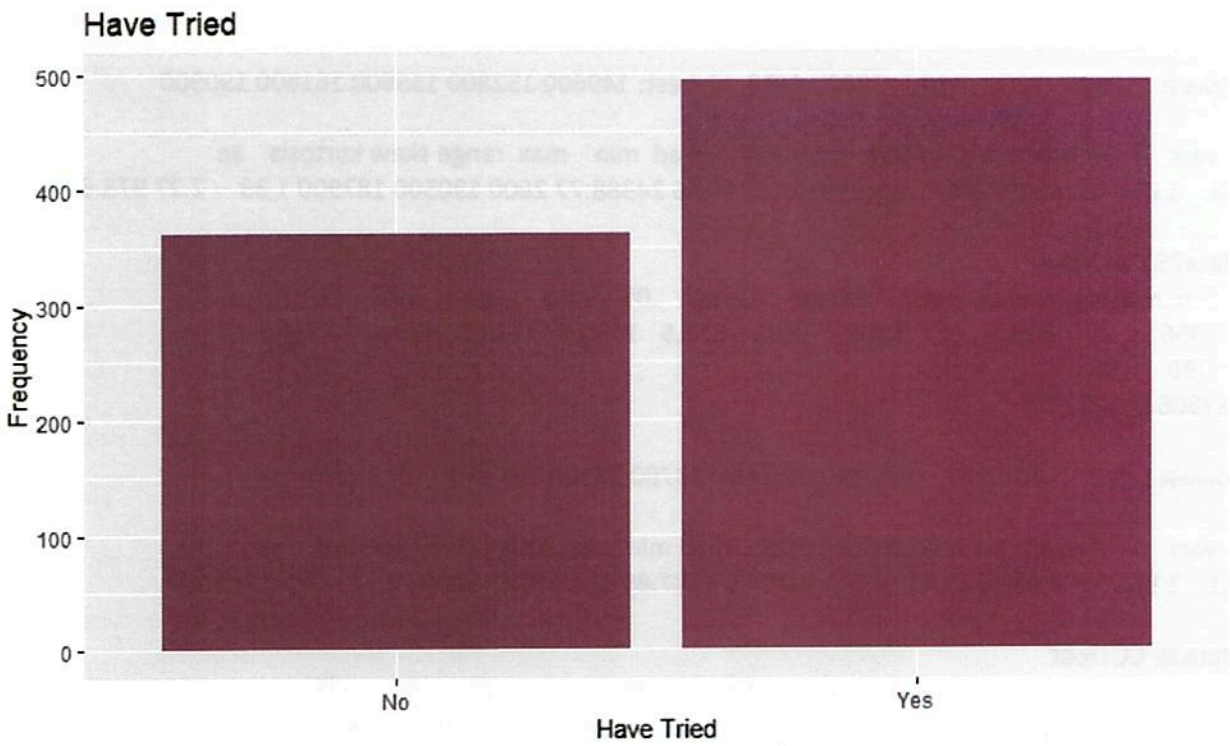

Pay Type
Frequency
500
400
300
200
100
0
Hourly
Salaried
Pay Type


Gender
400
300
200
100
0
Frequency
Female
Male
Gender


Live Alone
600
400
200
0
Frequency
No
Yes
Live Alone


Dwell Type
400
300
200
100
0
Frequency
Apt
Condo
Home
Dwell Type


Neighborhood
300
200
100
0
Frequency
East
South
West
Neighborhoods



data2$Age

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75
856	0	42	0.999	38.78	10.94	25.0	27.0	31.0	37.5	46.0

.90	.95
53.0	57.0

lowest : 22 23 24 25 26, highest: 59 60 61 62 64

vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	856	38.78	9.61	37.5	38.21	11.12	22	64	42	0.45	-0.69	0.33

data2$Weight

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75
856	0	112	1	192.7	28.21	155	162	174	190	210

.90	.95
228	237

lowest : 142 144 145 146 147, highest: 253 254 255 257 258

vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	856	192.66	24.75	190	191.79	25.2	142	258	116	0.3	-0.55	0.85

data2$Income

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75
856	0	546	1	45267	30438	9975	14850	24475	39950	58225

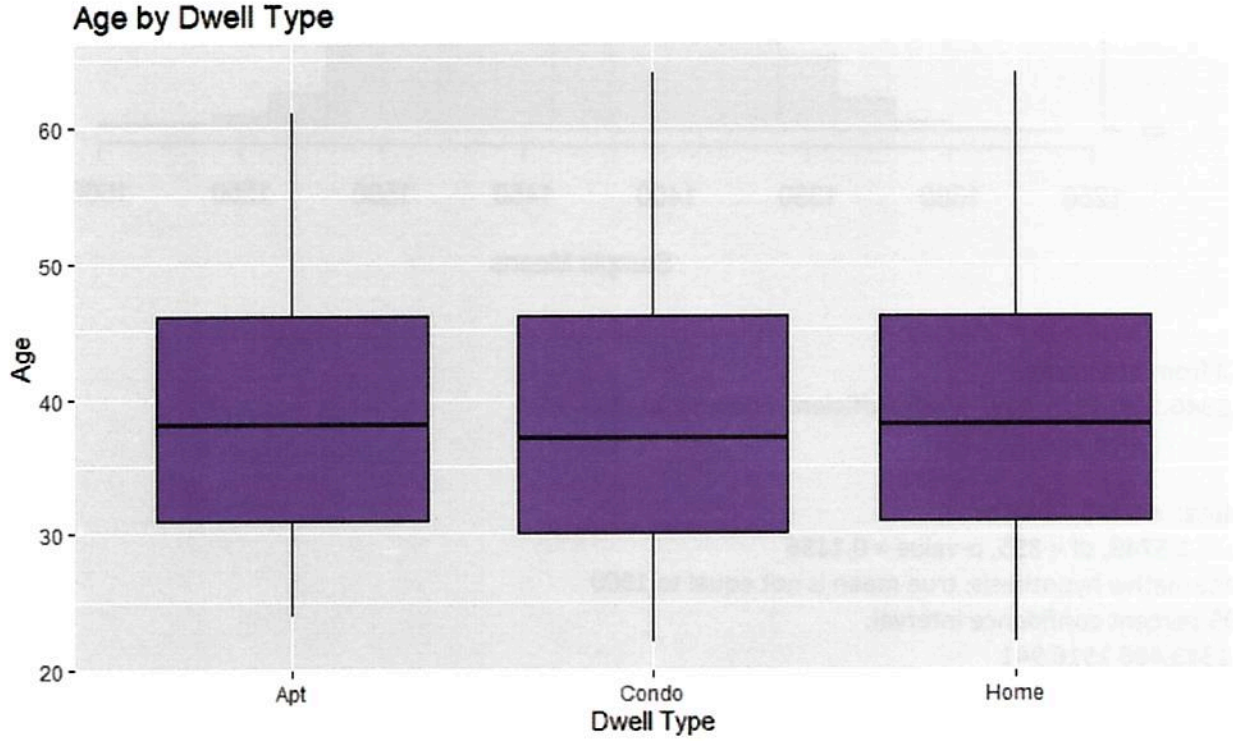
.90	.95

```
  80800  106950

lowest :  2600  3100  3200  3800  4400, highest: 149600 152800 155600 161600 190500

  vars   n    mean     sd median  trimmed     mad  min    max  range skew kurtosis   se
X1    1 856 45266.94 28631.29  39950 41744.46 24388.77 2600 190500 187900 1.33     2.27 978.6

data2$`Car Value`
     n  missing distinct    Info    Mean     Gmd     .05     .10     .25     .50     .75
   856        0      613       1    5908    5462   677.5  1090.0  2110.0  4175.0  7717.5
   .90     .95
 12605.0 17135.0

lowest :  130  150  210  280  290, highest: 29780 29910 29970 30710 33870

  vars   n    mean     sd median trimmed    mad min   max range skew kurtosis    se
X1    1 856 5908.48 5533.46   4175 4920.73 3587.89 130 33870 33740 1.97     4.53 189.13

data2$`CC Debt`
     n  missing distinct    Info    Mean     Gmd     .05     .10     .25     .50     .75
   856        0      337       1    1431    1294     130     265     560    1020    1972
   .90     .95
  3165    3785

lowest :    0   70   80   90  100, highest: 6970 7000 7460 8080 8960

  vars   n   mean     sd median trimmed    mad min  max range skew kurtosis   se
X1    1 856 1431.2 1278.04   1020 1232.46 889.56   0 8960  8960 1.84     4.56 43.68

data2$`Mall Trips`
     n  missing distinct    Info    Mean     Gmd     .05     .10     .25     .50     .75
   856        0       14   0.986   4.735   2.959       1       2       3       4       7
   .90     .95
     8      10

lowest :  0  1  2  3  4, highest:  9 10 11 12 17

Value       0    1    2    3    4    5    6    7    8    9   10   11   12   17
Frequency    23   57  100  134  121  122   83   86   49   32   26   21    1    1
Proportion 0.027 0.067 0.117 0.157 0.141 0.143 0.097 0.100 0.057 0.037 0.030 0.025 0.001 0.001

  vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 856 4.73 2.64      4    4.59 2.97   0  17    17 0.53     0.02 0.09
```

	East	South	West	Grand Total
Apt	69	50	94	213
Condo	80	57	102	239
Home	128	100	176	404
Grand Total	277	207	372	856

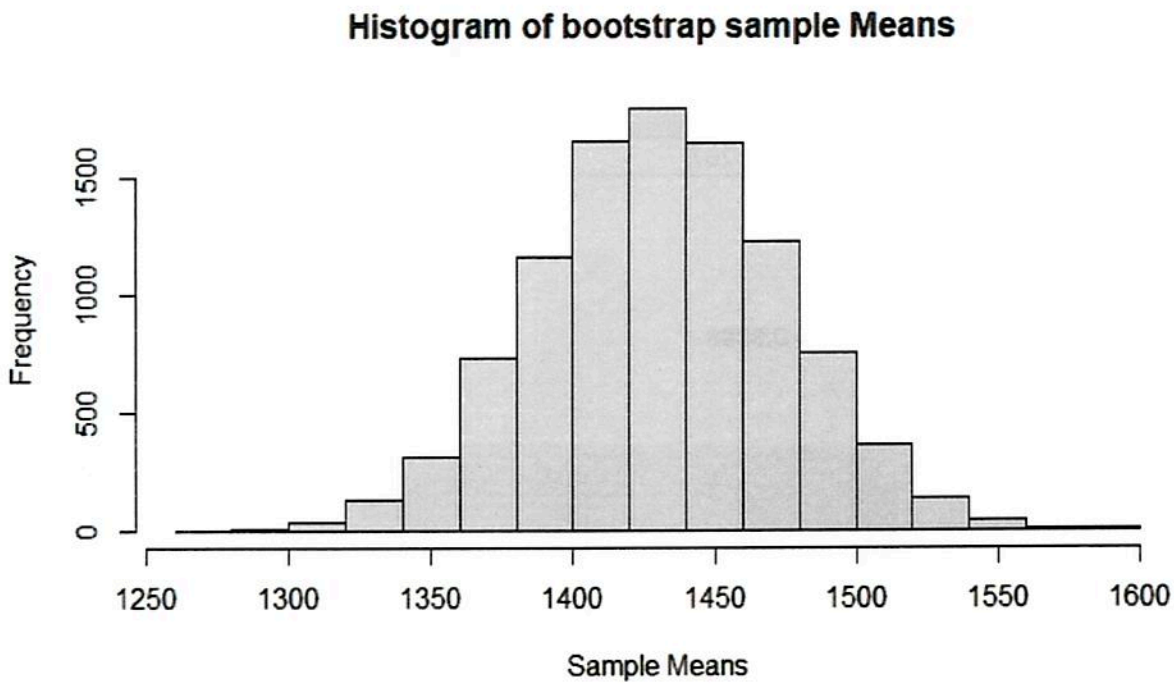
Pearson's Chi-squared test

data: observed_table
X-squared = 0.31492, df = 4, p-value = 0.9888



Df Sum Sq Mean Sq F value Pr(>F)
`Dwell Type` 2 32 16.24 0.175 0.839
Residuals 853 78941 92.55

No difference



CI from bootstrap:
(1346.519, 1515.409) -- not sufficient evidence to think so

One Sample t-test

data: data2$`CC Debt`
t = -1.5749, df = 855, p-value = 0.1156
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:
1345.466 1516.941
sample estimates:
mean of x
1431.203

(1345.466, 1516.941)

	No	Yes	Grand Total
Female	332	66	398
Male	379	79	458
Grand Total	711	145	856

2-sample test for equality of proportions with continuity correction

data: c out of c66 out of 39879 out of 458
X-squared = 0.028141, df = 1, p-value = 0.8668

alternative hypothesis: two.sided
95 percent confidence interval:
 -0.05933033 0.04601046
sample estimates:
  prop 1   prop 2
0.1658291 0.1724891

3.

Chi-squared test for given probabilities

data: observed
X-squared = 8.4848, df = 10, p-value = 0.5816

MTH 324, Final Exam, Fall 2022 Name KEY

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:
Use the work you did at home to answer these questions about tax paid and the neighborhoods in our dataset.

1. Based on the data from sheet 1 on ad preferences, is the data distributed normally or approximately so? Explain.

the differences look a little better, but all graphs have deviations from normal near the tails.

2. Report the results of your paired test of the ad preference data. Which test did you use and why? Clearly state your hypotheses, the conclusion in the context of the problem, and explain why you came to that conclusion.

$H_0: \delta = 0$
$H_a: \delta \neq 0$

p-value $4.16 \times 10^{-13} < 0.05$
reject H_0.

There is strong evidence that one ad is preferred over the other.

3. Based on the graphs of second data set, which of the numerical variables was most symmetric or normal?

weight

4. Based on the graphs of the second data set, which of the categorical variables had the biggest difference in frequency?

live alone yes vs. no

5. Based on the numerical summary of Income, provide the 5-number summary.

min: 2600
Q1: 24,475
median: 39,950
Q3: 58,225
max 190,500

6. Describe the results of your test of Independence for Dwell Type and Neighborhood. State your hypotheses, your conclusion in the context of the problem and explain your reasoning.

Ho: DwellType & Neighborhood are independent
Ha: " " " are not independent

p-value: 0.9888 > 0.05
fail to reject the null.
there is not sufficient evidence to think the variables are strongly dependent.

7. Describe the results of your ANOVA test for Dwell Type and Age. State your hypotheses, your conclusion in the context of the problem and explain your reasoning. Does your result agree with your boxplot?

Ho: $\mu_i = \mu_j$ for all i, j
Ha: $\mu_i \neq \mu_j$ for some $i \neq j$
p-value = 0.839
fail to reject null

there is not sufficient evidence that mean age differs by type of dwelling

8. Give your confidence interval from your bootstrap sample of Credit Card Debt. Give your confidence interval (95%) for the one-sample t- or z-interval. How do they compare?

bootsnap
(1346.5, 1515,4) your answers will vary slightly

from t-interval
(1345,5, 1516.9)

they are very similar

9. Describe the results of your test of two proportions for Live Alone by Gender. State your hypotheses, your conclusion in the context of the problem and explain your reasoning.

$H_0: p_1 = p_2$
$H_a: p_1 \neq p_2$

p-value: 0.8668

fail to reject H_0

there is not sufficient evidence to think the probability of living alone differs by gender

10. Based on the table in problem #3 on the at-home portion, describe the results of your goodness-of-fit test. State your hypotheses, your conclusion in the context of the problem and explain your reasoning.

H_0: data fits the distribution
H_a: data does not fit the distribution

p-value: 0.5816

fail to reject H_0

there is not sufficient evidence to think the data does not fit the distribution.

Part II:

11. If you needed to create a stratified sample in R (let's say on Gender), explain how you would go about doing that. (I don't need the code, but explain your steps in words.)

Separate the data into strata using filter
Select a random sample from each strata (a proportional amount based on strata size)
recombine into a single data set/sample.

answers may vary

12. A two-way table of Dwell Type and Neighborhood is shown below. Use it to answer the following questions.

	East	South	West	Grand Total
Apt	69	50	94	213
Condo	80	57	102	239
Home	128	100	176	404
Grand Total	277	207	372	856

a. What is the probability that a random person selected from this data set is from the East neighborhood?

$\frac{277}{856}$

b. What is the probability that a random person selected from this data set lives in a Condo?

$\frac{239}{856}$

c. What is the probability that a random person selected from this data set is from the East neighborhood and lives in a Condo?

$\frac{80}{856}$

d. What is the probability that a random person selected from this data set is from the East neighborhood or lives in a Condo?

$\frac{277+239-80}{856} = \frac{436}{856}$

e. What is the probability that a random person selected from this data set is from the East neighborhood given that they live in a condo?

$\frac{80}{239}$

f. Are the variables Neighborhood and Dwell Type independent or dependent? Does your answer differ if you consider only the descriptive properties of the table, or if you infer the answer from the hypothesis you conducted in Part 1? If it does, explain why.

$\frac{277}{856} = .323598...$ $\frac{80}{239} = 0.3347...$

these are not identical ($P(A) \neq P(A|B)$)

therefore, they are dependent.

13. The proportion of women in the sample on sheet 2 of the data set from the at-home portion of the exam is 0.465. If we were to randomly select 15 subjects from that data set, answer the following questions about the probability of possible outcomes.

a. What is the probability of getting exactly 8 women in the sample?

$\binom{15}{8}(0.465)^8(1-0.465)^7 = 0.17646...$

b. What is the probability of having fewer than 3 women in the sample?

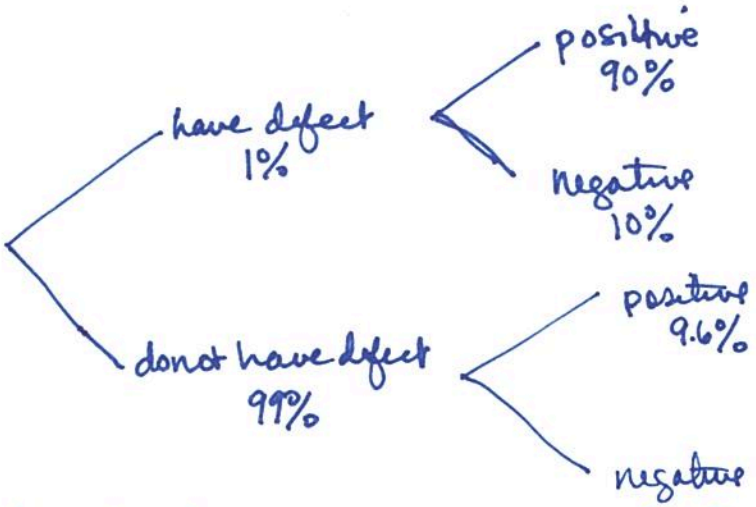
0, 1, 2

0.00786...

c. What is the expected number of women in the sample?

$15 \times .465 = 6.975$

14. Suppose that 1% of people have a certain genetic defect. Further suppose that 90% of tests for the gene detect the defect (true positives), and 9.6% of the tests are false positives. If a person gets a positive test result, what is the probability they actually have the genetic defect?



8.65%

$\frac{(0.01)(0.90)}{(0.01)(0.90) + (0.99)(0.096)} = 0.0865$

15. Consider the probability density function $f(x) = \frac{x^3}{5000}(10 - x), 0 \leq x \leq 10$ (it is equal to 0 everywhere else). Use this information to answer the questions that follow.

a. Verify that this function represents a valid probability distribution.

$$\int_0^{10} \frac{x^3}{5000}(10-x)\,dx = 1$$

b. Find $P(1 \leq X \leq 4)$

$$\int_1^4 \frac{x^3}{5000}(10-x)\,dx \approx 0.08658$$

c. Find the mean (expected value) of the distribution.

$$\int_0^{10} \frac{x^4}{5000}(10-x)\,dx = 6.\overline{6} = \frac{20}{3}$$