R Final Report

Ming Chen

June 27, 2019

Contents

1 Part 1

1.1 Report descriptive statistics of the data set obtained in (a).

The report on descriptive statistics is shown in table 1 and table 2.

Table 1: Descriptive Statistics 1

	Wr.Hnd	NW.Hnd	Pulse	Height	Age
Min.	13.00	12.50	35.00	152.00	16.92
1st Qu.	17.50	17.50	66.75	165.00	17.67
Median	18.50	18.50	72.00	170.60	18.58
Mean	18.80	18.73	74.02	172.50	20.43
3rd Qu.	20.00	20.00	80.00	180.00	20.17
Max.	23.20	23.50	104.00	200.00	70.42

Table 2: Descriptive Statistics 2

Sex	W.Hnd	Fold	Clap	Exer	Smoke	M.I
Female:84	Left:12	L on R :72	Left: 28	Freq:85	Heavy: 7	Imperial: 58
Male:84	Right:156	Neither: 8	Neither: 33	None:14	Never:134	Metric :110
		R on L :88	Right :107	Some:69	Occas: 13	
					Regul: 14	

1.2 Use boxplot to show the distributions of the height of male and female students.

The distributions of the height of male and female students are illustrated in figure 1.

1.3 Which numerical variables might have an influence on the student's pulse?

To indentify variables that have influence on students' pulse, we can first calculate the correlations between Pulse and independent variable, and then

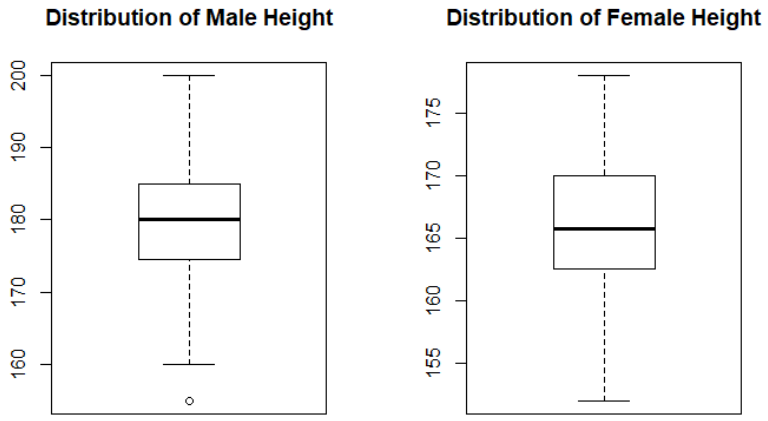

Figure 1: Distribution of Heights

conduct correlation test to make sure both variables are significantly correlated. Table 3 shows only "Clap" is significantly corelated with "Pulse" with a correlation of .198341. Hence we can conclude that "Clap" might be the factor that significantly influences pulse.

1.4 Is the probability of a student clapping his/her left hand on top less than 0.2?

To make sure whether the probability of a student clapping hand on top less than 0.2, we can use prop.test funtion to conduct proportion test. The null hypothesis is that "The probability of a student clapping his/her left hand on top greater than or equal to 0.2", and the alternative hypothesis is that "probability of a student clapping his/her left hand on top less than 0.2". The p-value is 0.1626 in this case, which suggests that we cannot reject the null hypothesis at .1 level of significance. Therefore we can infer that the probability of a student clapping his/her left hand on top is equal to or greater than 0.2.

Table 3: Correlations and P-Values

	cor(y,x)	P-value
Pulse and Age	-0.288	0.1144
Pulse and Height	-0.08326	0.283
Pulse and Wr.Hnd	-0.01382	0.859
Pulse and Sex	0.064165	0.409
Pulse and Sex	-0.04239	0.585
Pulse and Fold	-0.13951	0.071
Pulse and Exer	-0.05905	0.447
Pulse and Clap	0.198341	0.01
Pulse and W.Hnd	0.05421	0.485

1.5 Is the span of the writing hand in general larger than the span of the non-writing hand?

To make sure whether the span of the writing hand in general larger than the span of the non-writing hand, we can make a hypothesis test. The null hypothesis is 'the span of the writing hand is in general smaller than or equal to the span of the non-writing hand', and the alternative hypothesis is 'the span of the writing hand is in general larger to the span of the non-writing hand'. With a p-value of 0.3696, the hypothesis test result suggests that we cannot reject the null hypothesis. Therefore the the span of the writing hand is in general equal to or smaller than the span of the non-writing hand.

2 Part 2

2.1 According to CLT, what is the approximated distribution of the sample means?

According to CLT, the means of the sample tends towards a normal distribution as the number of samples gets greater. The approximated distribution of the sample means is normal distribution.

2.2 Draw the density plots of the sample means and its approximated distribution on one graph

The density plots are demonstrated in figure 2.

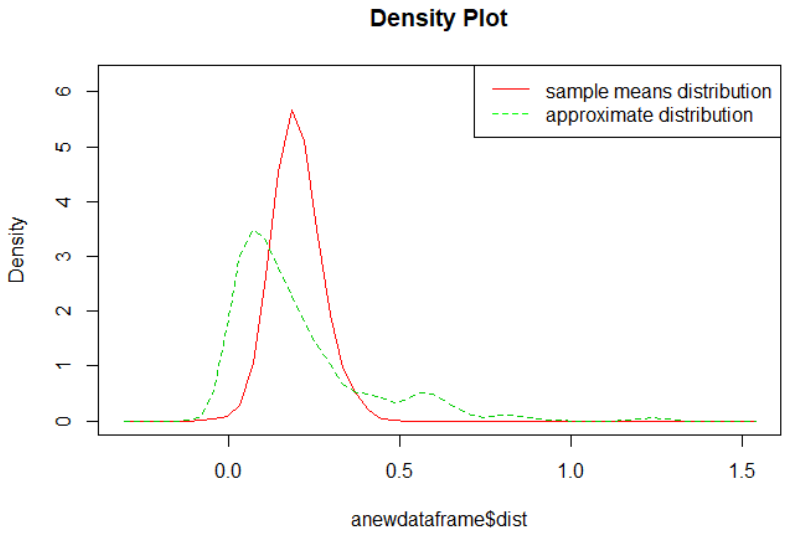

Figure 2: Density Plot

2.3 Show the qq plot of the distribution of the sample means

The QQ plot of the distribution of the sample means is demonstrated in figure 3.

2.4 What conclusion can you draw from this simulation?

From the density plot, we can know that the sample means distribution tends towards normal distribution as the sample means number gets greater. From QQ plot, we can also observe that the points fall on the straight 45-degree line, which suggests that the residual values are normally distributed with a mean of 0. Hence the sample means are normally distributed.

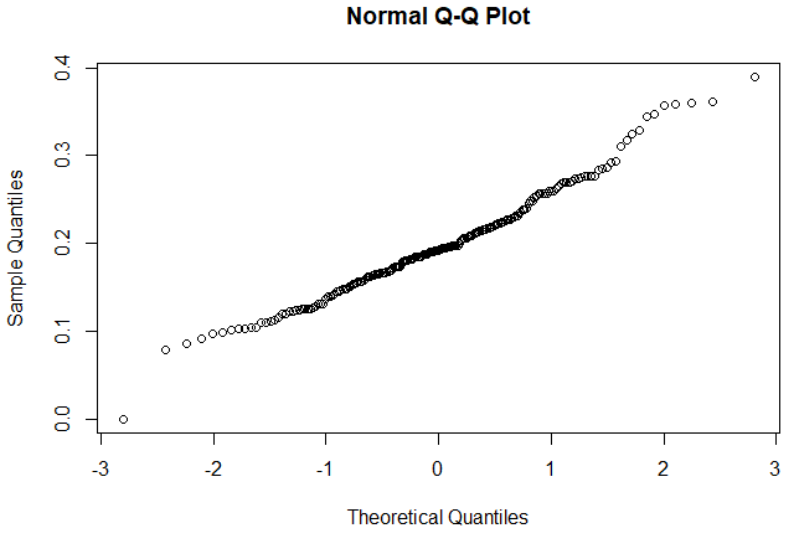

Figure 3: QQ Plot

3 Part 3

3.1 Draw 5000 random samples of X

The random samples generated are omitted here.

3.2 Given $\alpha = 0.05$, find the VaR of the samples obtained in 3.1

Since VaR denotes the critical value L such that $P(X > L) = \alpha P(X > L) = \alpha$,to find the VaR of the samples, we first need to sort the array of the change of the stock price ascendingly. Then we can use quantile function to find out the change value situated at given α. The VaR is -2.188176. (Note that as the values of samples vary each time we rerun the code, the VaR is unlikely to remain the current value)

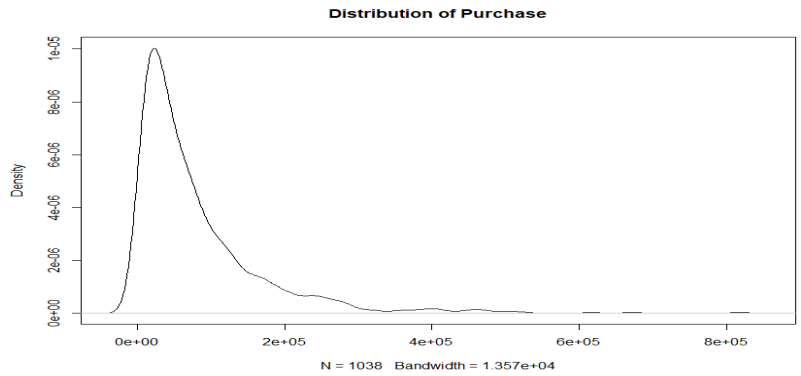

Figure 4: Distribution of Purchases Amount

3.3 Find the CVaR of the samples obtained in 3.1

CVaR is the average loss over a specified time period of unlikely scenarios beyond the confidence level. By calculating the arithmetic mean of rate of return under the critical value L, we can find the CVAR is -3.034234. (Note that as the values of samples vary each time we rerun the code, the CVaR is unlikely to remain the current value)

4 Part 4

4.1 Report on data exploration

We can use Distribution, pie charts, and boxplots, etc to explore on the data.

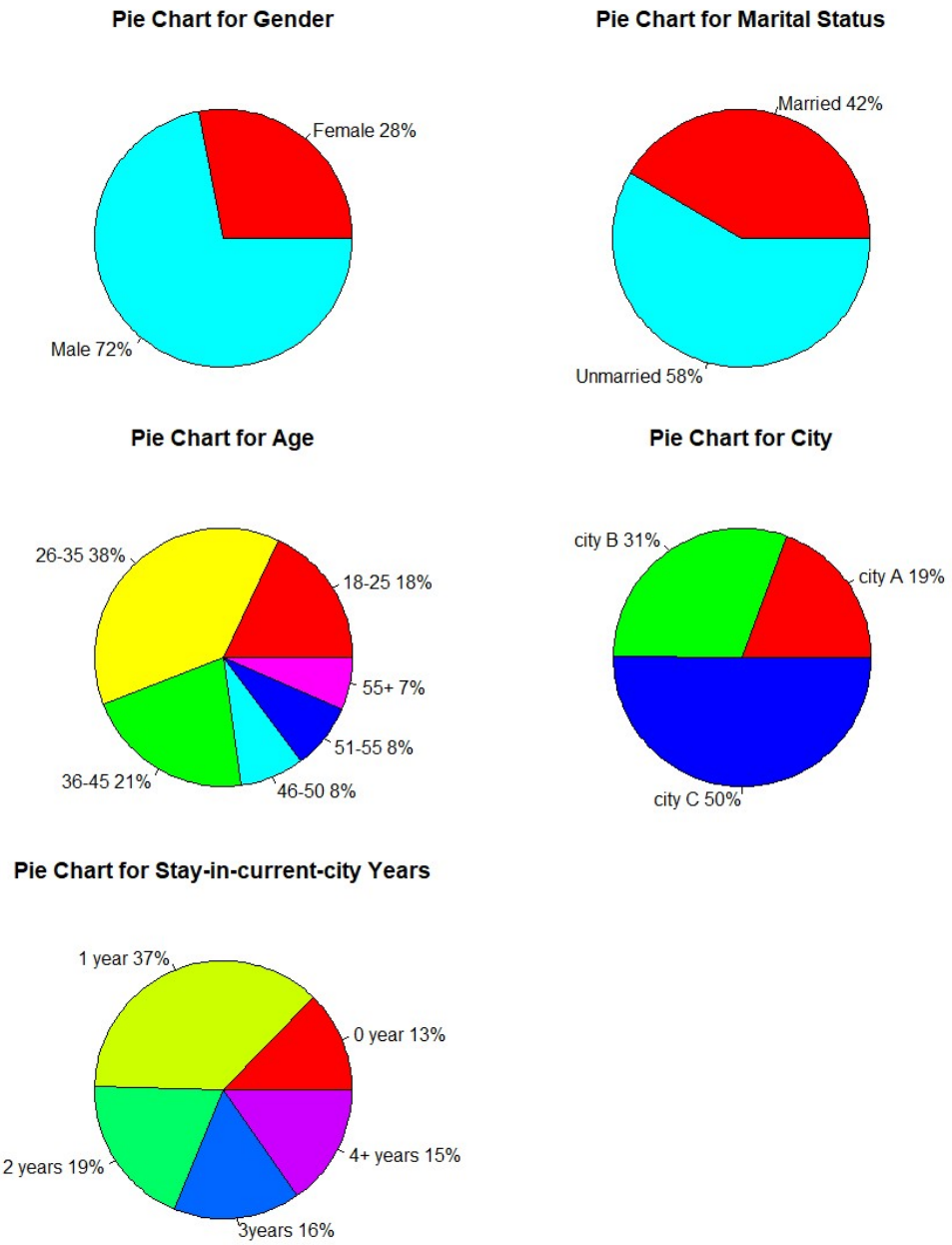

Figure 5: Pie Charts For Other Variables Distributions

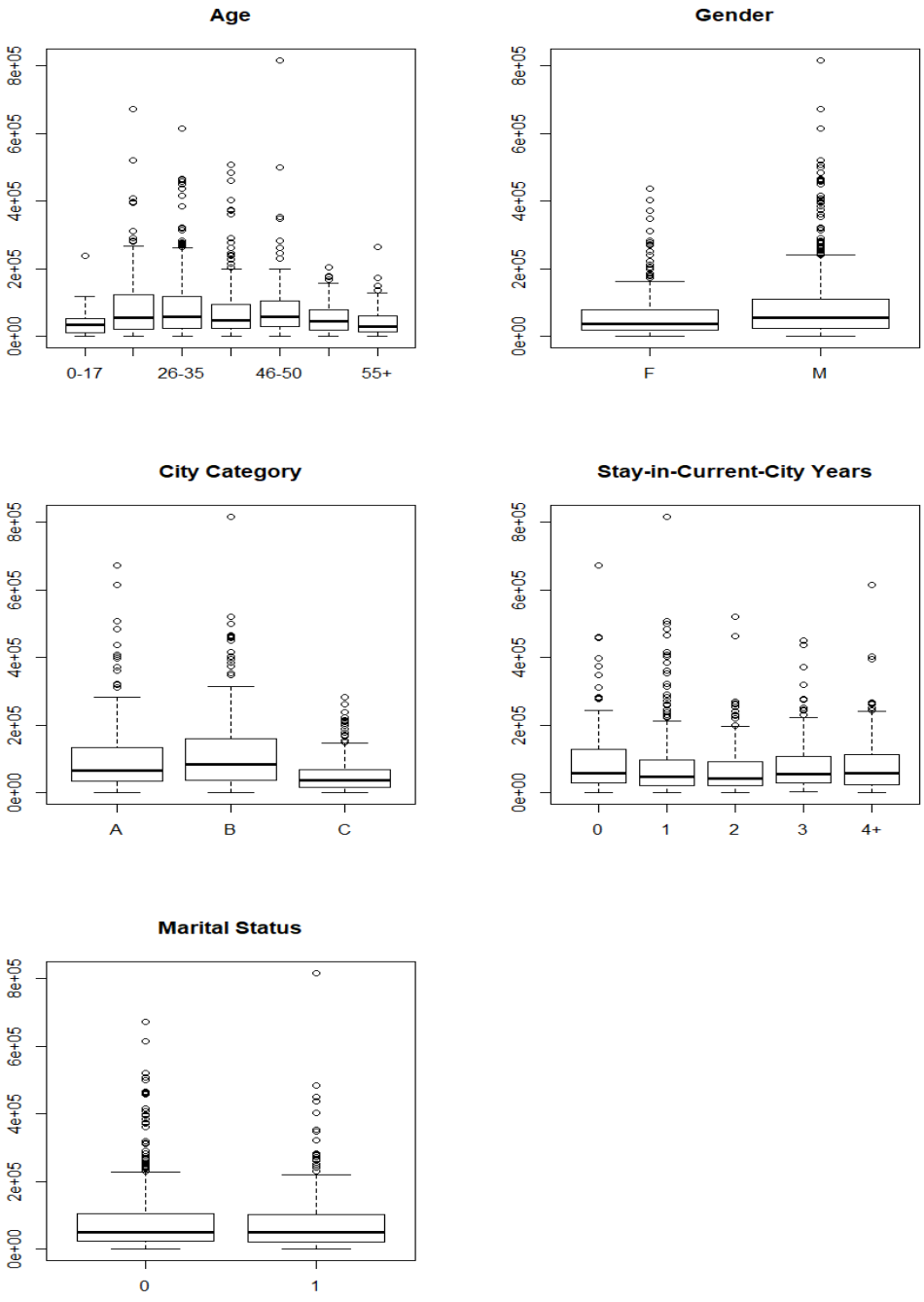

Figure 6: Box Plots For Other Variables Distributions

Table 4: Correlation Coefficients by Kendall Method

cor(y,x0)	cor(y,x1)	cor(y,x2)	cor(y,x3)	cor(y,x4)
0.0287828	-0.2512844	-0.10772	0.011602	-0.01244

4.2 Report on modeling (baseline model and alternatives)

We can take the independent varibles with the greatest correlation coefficient as the baseline model. Therefore, we begin constructing the regression model with 'city catergory' as the only independet variable. By adding independent values step by step in order of their correlation coefficients values. By adding independent variables step by step, we find that 'marital status' is not significantly related to purchases amount. In addition, we can also evaluate the model with "StepAIC" function. The result complies with the fitted model, lending supports for the validity of the fitted model. Hence we construct the fitted model with the remaining independent variables. The equation of the fitted model is:

$$y = 83820 - 55169x_1 + 25064x_2 + 37766x_3 - 20332x_4 - 27843x_5 \quad (1)$$

in which x_i are dummy varibles of certain factors.

4.3 Report on results and interpretation of the fitted model

Results of the fitted model have been demonstrated in table 5 and subsection 4.2. From table 5, we find that Purchases Amount is significantly related to 'City Category', 'Age', 'Gender', and 'Stay-in-current-city Years'. From the results we can conclude that the following factors significantly affect purchase amount: whether the customer live in city C, whether the customer is male, whether the customer is aged between 46 and 50, whether the customer stays in current city for 1 or 2 years. The fitted model implies that, when the aforementioned conditions are met, the customer is more likely to spend more on purchasing.

Table 5: P-values and Coefficients of The Fitted Regression Model

	$Pr(>\|t\|)$		Coefficients
(Intercept)	8.35E-08	***	83820
newdata$City_CategoryB	0.2375		8792
newdata$City_CategoryC	6.07E-15	***	-55169
newdata$GenderM	1.35E-05	***	25064
newdata$Age18-25	0.0887	.	24030
newdata$Age26-35	0.0654	.	24692
newdata$Age36-45	0.1004	.	22770
newdata$Age46-50	0.0161	*	37766
newdata$Age51-55	0.8677		2593
newdata$Age55+	0.8107		3890
newdata$Stay_In_Current_City_Years1	0.0154	*	-20332
newdata$Stay_In_Current_City_Years2	0.0028	**	-27843
newdata$Stay_In_Current_City_Years3	0.0728	.	-17374
newdata$Stay_In_Current_City_Years4+	0.114		-15446

4.4 Report on model assumptions

In Q-Q graph, points basically cluster around the line, which proves that the assumption of Normality is satisfied; there is no reason to assume any of the independents are correlated. Therefore, the assumption of Independence is satisfied; from graph one, we can observe that residuals have no systematic relationship between residuals and the predicted values. In the Residuals vs. Fitted graph, you see clear evidence of a straight relationship, and hence the model well captures systematic variance in the data, thereby proves that the assumption of Linearity is satisfied; the Scale-Location graph shows that the points form a random band around the horizontal line. Hence the assumption of Homoscedasticity is satisfied.

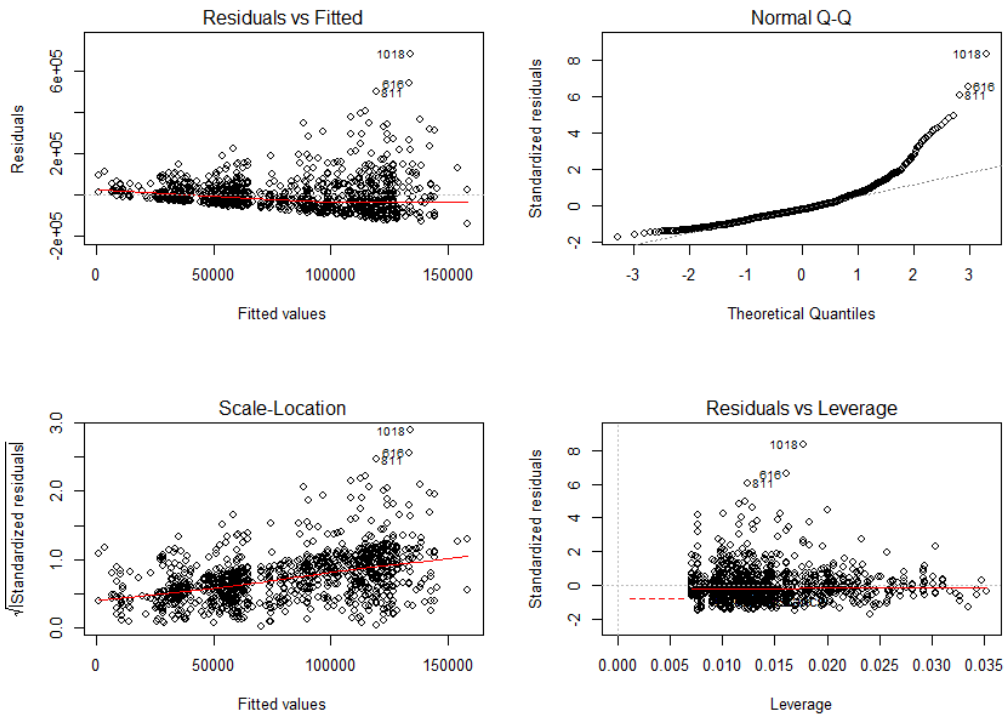

Figure 7: Diagnostic plots