**Social Networking Data**

The file ‘Pilot\_SN\_Data.csv’ is an outcome of a pilot-study survey on the usage of social networking among people in Kuwait. The description of the questionnaire items and variable and category coding is given in the file ‘Pilot\_SN\_Survey.pdf’. Note that the codes used for the categories of each question is marked in Blue. For example, the first variable (q01== Gender; 1== Male; 2== Female).

1. Rename the variable names with suitable text description for the variables q01-q05, q06\_01-q06-06, q08 & q09.

|  |
| --- |
| *R code:*  **Pilot <-read.csv(file.choose())**  **View(Pilot)**  **names(Pilot)[names(Pilot) == "q01"] <- "Gender"**  **names(Pilot)[names(Pilot) == "q02"] <- "Age"**  **names(Pilot)[names(Pilot) == "q03"] <- "Status In Kuwait"**  **names(Pilot)[names(Pilot) == "q04"] <- "Education"**  **names(Pilot)[names(Pilot) == "q05"] <- "Occupation"**  **names(Pilot)[names(Pilot) == "q06\_1"] <- "Twitter"**  **names(Pilot)[names(Pilot) == "q06\_02"] <- "Facebook"**  **names(Pilot)[names(Pilot) == "q06\_03"] <- "Youtube"**  **names(Pilot)[names(Pilot) == "q06\_04"] <- "Instagram"**  **names(Pilot)[names(Pilot) == "q06\_05"] <- "Whatsapp"**  **names(Pilot)[names(Pilot) == "q06\_06"] <- "LinkedIn"**  **names(Pilot)[names(Pilot) == "q08"] <- "Business IG"**  **names(Pilot)[names(Pilot) == "q09"] <- "Buy from IG"** |

|  |
| --- |
| *Output and Results:*  *Table, Excel  Description automatically generated* |

1. Perform a barplot for the variable **Age** (q02) with counts displayed on the Y-axis and another graph with percentages instead of counts. Rename the variable labels with suitable text description for categories as indicated in Blue in the questionnaire (i.e, Under 17== 1, 18-24 == 2, etc).

Complete the following statement:

* More than \_\_\_\_% of the participant are above the age of 34 yrs.

|  |
| --- |
| *R code:*  **counts <- table(Pilot$Age)**  **names(counts)[names(counts) == "1"] <- "Under 17"**  **names(counts)[names(counts) == "2"] <- "18-24"**  **names(counts)[names(counts) == "3"] <- "25-34"**  **names(counts)[names(counts) == "4"] <- "35-44"**  **names(counts)[names(counts) == "5"] <- "45-54"**  **names(counts)[names(counts) == "6"] <- "55 and older"**  **names(counts)[names(counts) == "7"] <- "Unknown"**  **names(counts)[names(counts) == "8"] <- "Unknown"**  **barplot(counts, main = "Participants by Age", xlab = "Age Group", ylab = "Count", col = "Blue")** |

|  |
| --- |
| *Output and Results:*  *Chart, bar chart  Description automatically generated* |

1. In the data set file, you will find in the last column a variable ‘AgeR’ that is not part of the questionnaire but it is a recoding of the variable ‘Age’. Do a frequency table to variable ‘AgeR’. Which age categories are being combined?

|  |
| --- |
| *R code:*  cross\_tab = xtabs(~ Age + AgeR, Pilot)  cross\_tab |

|  |
| --- |
| *Output and Results:*    *We can conclude that,*   * *category 1 & 2 from age are combined, to form category 1 in AgeR.* * *category 3, 4 & 5 from age are combined, to form category 1 in AgeR.* * *category 6, 7 & 8 from age are combined, to form category 1 in AgeR.* |

1. Repeat Q2 above for q06 from the questionnaire on the utilization of Social Networking (SN). Rename the variable labels with suitable text description for categories as indicated in Blue in the questionnaire.

Which of the six SN (Twitter, Facebook, YouTube, Instagram, WhatsApp, LinkedIn) has the highest percentage of usage?

|  |
| --- |
| *R code:*  value <- c(table(Pilot$Twitter)[1], table(Pilot$Facebook)[1], table(Pilot$Youtube)[1], table(Pilot$Instagram)[1], table(Pilot$Whatsapp)[1],table(Pilot$LinkedIn)[1])  names(value) <- c("Twitter", "Facebook", "Youtube", "Instagram", "Whatsapp", "LinkedIn")  barplot(value, main = "Social Media Usage", xlab = "Social Network", ylab = "Usage", col = "Red") |

|  |
| --- |
| *Output and Results:* |

1. Is there a relationship between ‘**Age’** and ‘**Buy form Instagram** **(q09)’**? (Use chi-square test of independence to answer this question. Make sure to

State clearly the

* null hypothesis and alternative hypothesis
* test statistic and its value
* decision and the conclusion in words

|  |
| --- |
| *R code:*  test <- chisq.test(table(Pilot$Age, Pilot$`Buy from IG`))  test  test$statistic  test$p.value |

|  |
| --- |
| *Output and Results:*  *Null hypothesis: Buy from IG is not dependent of age*  *Alternative hypothesis : Buy from IG is dependent on age*    *X-Squared value : 15.487*  *p-value = 0.03*  *Since, the p-value lies below the significance level of 0.05, hence the null hypothesis is rejected. Thus, the purchase from IG is dependent on the age* |

1. If you have a warning message with the chi-square analysis. Indicate why is that? What condition(s) are not satisfied?

|  |
| --- |
| *Output and Results:*  *The warning message indicates that, “the Chi-squared approximation maybe incorrect”. This occurs because the values to be expected are very small, since chi squared test is an approximation over normally distributed data.*    *In this case, the rows 7 and 8 could be the contributor to the problemsince their values are considerably smaller as compared to other age categories and could be merged to get rid of the warning.* |

1. Repeat the chi-square analysis as you did in exam Question 7 above with replacing ‘Age’ with the recoded age **‘AgeR’**. Does the result improve?

|  |
| --- |
| *R code:*  test <- chisq.test(table(Pilot$AgeR, Pilot$`Buy from IG`))  test  test$statistic  test$p.value  table(Pilot$AgeR, Pilot$`Buy from IG`)  test$observed  test$expected |

|  |
| --- |
| *Output and Results:*    *Observed and expected value for AgeR, without the warning sign and better approximation*  > test$observed    1 2  1 95 43  2 35 19  3 16 20  > test$expected    1 2  1 88.36842 49.63158  2 34.57895 19.42105  3 23.05263 12.94737  *Observed and expected value for AgeR, with the warning sign*  > test$observed    1 2  1 44 16  2 51 27  3 13 12  4 12 6  5 10 1  6 11 12  7 4 8  8 1 0  > test$expected    1 2  1 38.4210526 21.5789474  2 49.9473684 28.0526316  3 16.0087719 8.9912281  4 11.5263158 6.4736842  5 7.0438596 3.9561404  6 14.7280702 8.2719298  7 7.6842105 4.3157895  8 0.6403509 0.3596491  *Yes, the results have improved after replacing age with ageR* |

1. Do logistic regression to predict **‘Having business Instagram account (q08)’** from all demographic variables (Gender, AgeR, citizenship, Education, Occupation) in addition to q09 (buying form Instagram). Comment on your model and summarize your findings in a paragraph of at least 5 lines.

|  |
| --- |
| *R code:*    #convert columns to factors for logistic regression  cat <- c("Business\_IG", "Buy\_from\_IG", "Gender", "AgeR", "Status\_In\_Kuwait", "Education", "Occupation")  Pilot[cat] <- lapply(Pilot[cat], factor)  #logistic regression  data = glm(formula = Business\_IG ~ Buy\_from\_IG + Gender + AgeR + Status\_In\_Kuwait + Education + Occupation, data = Pilot, family = binomial)  print(summary(data)) |

|  |
| --- |
| *Output and Results:*  *The deciding columns need to be converted into categories i.e factors, because logistic regression performs classification based on categorical values.*  > print(summary(data))  Call:  glm(formula = Business\_IG ~ Buy\_from\_IG + Gender + AgeR + Status\_In\_Kuwait +  Education + Occupation, family = binomial, data = Pilot)  Deviance Residuals:  Min 1Q Median 3Q Max  -2.4086 0.3363 0.5069 0.5831 1.1382  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) 1.5095 1.3399 1.127 0.2599  Buy\_from\_IG2 0.3490 0.4907 0.711 0.4770  Gender2 -0.3759 0.5852 -0.642 0.5206  AgeR2 0.3669 0.6914 0.531 0.5956  AgeR3 0.1125 0.8393 0.134 0.8933  Status\_In\_Kuwait2 -0.8591 0.8714 -0.986 0.3242  Status\_In\_Kuwait3 -1.8940 1.0729 -1.765 0.0775 .  Education2 1.7107 1.3821 1.238 0.2158  Education3 0.7281 1.3234 0.550 0.5822  Education4 0.8535 1.2741 0.670 0.5029  Education5 0.6015 1.7123 0.351 0.7254  Occupation2 -0.7624 0.6961 -1.095 0.2734  Occupation3 -0.7563 1.2069 -0.627 0.5309  Occupation4 -0.3050 0.9109 -0.335 0.7378  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 177.56 on 227 degrees of freedom  Residual deviance: 168.06 on 214 degrees of freedom  (3 observations deleted due to missingness)  AIC: 196.06  Number of Fisher Scoring iterations: 5  *In the logistic regression, the family chosen is binomial because the outcome of “having a business IG” has only 2 outcomes 1 or 2.*  *The summary prints the statistics of the logistic regression, to identify the outcome.*  *Here, we can see that the p-values of all the columns are greater than 0.05 significance level. This means that none of the chosen columns effect having a business account on IG.* |