assignment 5 part 2

Kylia Medeiros

2024-10-17

library(readxl)

## Warning: package 'readxl' was built under R version 4.3.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tibble' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

df <- read\_excel('c2015.xlsx')

-2.)Let’s study the variable SEX. How many missing values in the NAs form?

colSums(df == 'SEX', na.rm = TRUE)

## STATE ST\_CASE VEH\_NO PER\_NO COUNTY DAY MONTH HOUR   
## 0 0 0 0 0 0 0 0   
## MINUTE AGE SEX PER\_TYP INJ\_SEV SEAT\_POS DRINKING YEAR   
## 0 0 0 0 0 0 0 0   
## MAN\_COLL OWNER MOD\_YEAR TRAV\_SP DEFORMED DAY\_WEEK ROUTE LATITUDE   
## 0 0 0 0 0 0 0 0   
## LONGITUD HARM\_EV LGT\_COND WEATHER   
## 0 0 0 0

-3.)Still with variable SEX. There are missing values in this variables that are not NAs. Identify the forms of missing values in this variable. Change all the forms of missing values to NAs.

df <- replace(df, df == 'Unknown' |  
 df == 'Not Rep' |  
 df == 'Not Available', NA)

-4.)Still with variable SEX. After all the missing values are in the NAs form. Change the missing values of this variable to the majority sex.

majority\_sex <-names(which.max(table(df$SEX)))  
df$SEX <- replace\_na(df$SEX, majority\_sex)  
table(df$SEX)

##   
## Female Male   
## 26631 53956

5.) Let’s study variable AGE. Use the table function to check out the values of these variable and forms of missing values. Use na\_if to change all the forms of missing values to NAs.

df$AGE <- na\_if(df$AGE, "")  
df$AGE <- na\_if(df$AGE, "N/A")  
table(df$AGE, useNA = "ifany")

##   
## 1 10 101 103 11 114   
## 441 378 1 1 369 3   
## 12 13 14 15 16 17   
## 401 414 525 718 1091 1530   
## 18 19 2 20 21 22   
## 1899 1939 410 1931 2029 2051   
## 23 24 25 26 27 28   
## 1994 1893 1820 1676 1613 1374   
## 29 3 30 31 32 33   
## 1495 418 1340 1366 1298 1232   
## 34 35 36 37 38 39   
## 1228 1256 1126 1184 1102 1046   
## 4 40 41 42 43 44   
## 374 1036 1043 1048 1003 1120   
## 45 46 47 48 49 5   
## 1072 1071 1045 1033 1068 389   
## 50 51 52 53 54 55   
## 1147 1231 1128 1175 1172 1155   
## 56 57 58 59 6 60   
## 1095 1080 1003 924 395 895   
## 61 62 63 64 65 66   
## 811 806 760 722 679 674   
## 67 68 69 7 70 71   
## 625 701 577 361 521 492   
## 72 73 74 75 76 77   
## 530 400 382 401 376 330   
## 78 79 8 80 81 82   
## 323 294 384 313 261 274   
## 83 84 85 86 87 88   
## 262 249 237 202 187 168   
## 89 9 90 91 92 93   
## 138 426 109 87 67 51   
## 94 95 96 97 98 99   
## 45 27 16 12 6 2   
## Less than 1 <NA>   
## 337 1668

-6.) Still with variable AGE. Use the str\_replace to replace Less than 1 to ‘0’ (character 0, not number 0).

df$AGE <- str\_replace(df$AGE, "Less than 1", "0")  
table(df$AGE, useNA = "ifany")

##   
## 0 1 10 101 103 11 114 12 13 14 15 16 17 18 19 2   
## 337 441 378 1 1 369 3 401 414 525 718 1091 1530 1899 1939 410   
## 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34   
## 1931 2029 2051 1994 1893 1820 1676 1613 1374 1495 418 1340 1366 1298 1232 1228   
## 35 36 37 38 39 4 40 41 42 43 44 45 46 47 48 49   
## 1256 1126 1184 1102 1046 374 1036 1043 1048 1003 1120 1072 1071 1045 1033 1068   
## 5 50 51 52 53 54 55 56 57 58 59 6 60 61 62 63   
## 389 1147 1231 1128 1175 1172 1155 1095 1080 1003 924 395 895 811 806 760   
## 64 65 66 67 68 69 7 70 71 72 73 74 75 76 77 78   
## 722 679 674 625 701 577 361 521 492 530 400 382 401 376 330 323   
## 79 8 80 81 82 83 84 85 86 87 88 89 9 90 91 92   
## 294 384 313 261 274 262 249 237 202 187 168 138 426 109 87 67   
## 93 94 95 96 97 98 99 <NA>   
## 51 45 27 16 12 6 2 1668

-7.)Still with variable AGE. Use the class function to check the type of this variable. Use the as.numeric function to change the type of the variable to numeric.

class(df$AGE)

## [1] "character"

df$AGE <- as.numeric(df$AGE)  
class(df$AGE)

## [1] "numeric"

-8.)Still with variable AGE. Replace the missing values NAs by the mean of the variable.

mean\_age <- mean(df$AGE, na.rm = TRUE)  
df$AGE[is.na(df$AGE)] <- mean\_age  
sum(is.na(df$AGE))

## [1] 0

1. Let’s fix the variableTRAV\_SP`. Do the follows.

Use the table function to check all the values of this variable. Use the str\_remove to remove the MPH in each value.

table(df$TRAV\_SP, useNA = "ifany")

##   
## 001 MPH 002 MPH 003 MPH 004 MPH 005 MPH 006 MPH 007 MPH 008 MPH 009 MPH 010 MPH   
## 36 36 34 13 632 20 30 41 11 828   
## 011 MPH 012 MPH 013 MPH 014 MPH 015 MPH 016 MPH 017 MPH 018 MPH 019 MPH 020 MPH   
## 25 26 26 25 706 17 26 34 15 516   
## 021 MPH 022 MPH 023 MPH 024 MPH 025 MPH 026 MPH 027 MPH 028 MPH 029 MPH 030 MPH   
## 16 29 31 19 616 13 20 53 14 789   
## 031 MPH 032 MPH 033 MPH 034 MPH 035 MPH 036 MPH 037 MPH 038 MPH 039 MPH 040 MPH   
## 23 38 71 35 1446 25 41 122 45 1430   
## 041 MPH 042 MPH 043 MPH 044 MPH 045 MPH 046 MPH 047 MPH 048 MPH 049 MPH 050 MPH   
## 48 47 135 33 2708 44 47 143 81 1643   
## 051 MPH 052 MPH 053 MPH 054 MPH 055 MPH 056 MPH 057 MPH 058 MPH 059 MPH 060 MPH   
## 44 65 189 44 3691 50 105 185 68 1517   
## 061 MPH 062 MPH 063 MPH 064 MPH 065 MPH 066 MPH 067 MPH 068 MPH 069 MPH 070 MPH   
## 68 98 143 47 1912 67 68 243 31 1614   
## 071 MPH 072 MPH 073 MPH 074 MPH 075 MPH 076 MPH 077 MPH 078 MPH 079 MPH 080 MPH   
## 76 61 113 24 613 43 38 88 40 527   
## 081 MPH 082 MPH 083 MPH 084 MPH 085 MPH 086 MPH 087 MPH 088 MPH 089 MPH 090 MPH   
## 38 29 44 23 205 30 15 34 15 209   
## 091 MPH 092 MPH 093 MPH 094 MPH 095 MPH 096 MPH 097 MPH 098 MPH 099 MPH 100 MPH   
## 41 13 15 10 53 23 36 11 16 187   
## 101 MPH 102 MPH 103 MPH 104 MPH 105 MPH 106 MPH 107 MPH 108 MPH 109 MPH 110 MPH   
## 9 5 11 8 17 12 8 3 6 24   
## 111 MPH 112 MPH 113 MPH 114 MPH 115 MPH 116 MPH 118 MPH 119 MPH 120 MPH 122 MPH   
## 3 15 6 5 13 5 5 9 22 1   
## 123 MPH 124 MPH 125 MPH 126 MPH 128 MPH 130 MPH 134 MPH 137 MPH 140 MPH 150 MPH   
## 2 7 1 4 2 7 1 1 3 2   
## Greater Stopped <NA>   
## 1 3129 51419

df$TRAV\_SP <- str\_remove(df$TRAV\_SP, "MPH")  
table(df$TRAV\_SP, useNA = "ifany")

##   
## 001 002 003 004 005 006 007 008 009 010   
## 36 36 34 13 632 20 30 41 11 828   
## 011 012 013 014 015 016 017 018 019 020   
## 25 26 26 25 706 17 26 34 15 516   
## 021 022 023 024 025 026 027 028 029 030   
## 16 29 31 19 616 13 20 53 14 789   
## 031 032 033 034 035 036 037 038 039 040   
## 23 38 71 35 1446 25 41 122 45 1430   
## 041 042 043 044 045 046 047 048 049 050   
## 48 47 135 33 2708 44 47 143 81 1643   
## 051 052 053 054 055 056 057 058 059 060   
## 44 65 189 44 3691 50 105 185 68 1517   
## 061 062 063 064 065 066 067 068 069 070   
## 68 98 143 47 1912 67 68 243 31 1614   
## 071 072 073 074 075 076 077 078 079 080   
## 76 61 113 24 613 43 38 88 40 527   
## 081 082 083 084 085 086 087 088 089 090   
## 38 29 44 23 205 30 15 34 15 209   
## 091 092 093 094 095 096 097 098 099 100   
## 41 13 15 10 53 23 36 11 16 187   
## 101 102 103 104 105 106 107 108 109 110   
## 9 5 11 8 17 12 8 3 6 24   
## 111 112 113 114 115 116 118 119 120 122   
## 3 15 6 5 13 5 5 9 22 1   
## 123 124 125 126 128 130 134 137 140 150   
## 2 7 1 4 2 7 1 1 3 2   
## Greater Stopped <NA>   
## 1 3129 51419

df <- df %>%  
 filter(TRAV\_SP != "Greater") %>%  
 drop\_na(TRAV\_SP)  
table(df$TRAV\_SP, useNA = "ifany")

##   
## 001 002 003 004 005 006 007 008 009 010   
## 36 36 34 13 632 20 30 41 11 828   
## 011 012 013 014 015 016 017 018 019 020   
## 25 26 26 25 706 17 26 34 15 516   
## 021 022 023 024 025 026 027 028 029 030   
## 16 29 31 19 616 13 20 53 14 789   
## 031 032 033 034 035 036 037 038 039 040   
## 23 38 71 35 1446 25 41 122 45 1430   
## 041 042 043 044 045 046 047 048 049 050   
## 48 47 135 33 2708 44 47 143 81 1643   
## 051 052 053 054 055 056 057 058 059 060   
## 44 65 189 44 3691 50 105 185 68 1517   
## 061 062 063 064 065 066 067 068 069 070   
## 68 98 143 47 1912 67 68 243 31 1614   
## 071 072 073 074 075 076 077 078 079 080   
## 76 61 113 24 613 43 38 88 40 527   
## 081 082 083 084 085 086 087 088 089 090   
## 38 29 44 23 205 30 15 34 15 209   
## 091 092 093 094 095 096 097 098 099 100   
## 41 13 15 10 53 23 36 11 16 187   
## 101 102 103 104 105 106 107 108 109 110   
## 9 5 11 8 17 12 8 3 6 24   
## 111 112 113 114 115 116 118 119 120 122   
## 3 15 6 5 13 5 5 9 22 1   
## 123 124 125 126 128 130 134 137 140 150   
## 2 7 1 4 2 7 1 1 3 2   
## Stopped   
## 3129

df$TRAV\_SP <- str\_replace(df$TRAV\_SP, "Stopped", "0")  
table(df$TRAV\_SP, useNA = "ifany")

##   
## 0 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015   
## 3129 36 36 34 13 632 20 30 41 11 828 25 26 26 25 706   
## 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031   
## 17 26 34 15 516 16 29 31 19 616 13 20 53 14 789 23   
## 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047   
## 38 71 35 1446 25 41 122 45 1430 48 47 135 33 2708 44 47   
## 048 049 050 051 052 053 054 055 056 057 058 059 060 061 062 063   
## 143 81 1643 44 65 189 44 3691 50 105 185 68 1517 68 98 143   
## 064 065 066 067 068 069 070 071 072 073 074 075 076 077 078 079   
## 47 1912 67 68 243 31 1614 76 61 113 24 613 43 38 88 40   
## 080 081 082 083 084 085 086 087 088 089 090 091 092 093 094 095   
## 527 38 29 44 23 205 30 15 34 15 209 41 13 15 10 53   
## 096 097 098 099 100 101 102 103 104 105 106 107 108 109 110 111   
## 23 36 11 16 187 9 5 11 8 17 12 8 3 6 24 3   
## 112 113 114 115 116 118 119 120 122 123 124 125 126 128 130 134   
## 15 6 5 13 5 5 9 22 1 2 7 1 4 2 7 1   
## 137 140 150   
## 1 3 2

df$TRAV\_SP <- na\_if(df$TRAV\_SP, "")  
df$TRAV\_SP <- na\_if(df$TRAV\_SP, "N/A")  
sum(is.na(df$TRAV\_SP))

## [1] 0

df$TRAV\_SP <- as.numeric(df$TRAV\_SP)  
class(df$TRAV\_SP)

## [1] "numeric"

median\_trav\_sp <- median(df$TRAV\_SP, na.rm = TRUE)  
df$TRAV\_SP <- replace\_na(df$TRAV\_SP, median\_trav\_sp)

-10.)Find the correlation between Age of the drivers and Travel speed (TRAV\_SP). Hint: You want to look at the seat positions (SEAT\_POS variable) to filter out the observations about the drivers, then calculate the correlation.

drivers\_data <- df %>% filter(SEAT\_POS == "Driver")