**Source code**

**Data Quality Check / Data Cleaning**

# Tell R where the 3 datasets are located

accidents <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/accidents.csv")

casualties <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/casualties.csv")

weathers <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/weather.csv")

vehicles <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/vehicles.csv")

View(accidents)

View(casualties)

View(vehicles)

rm(casualties1)

# Delete unwanted observation (years)

accident <- accidents[-c(646831:738029), ]

#casualties\_df <- casualties[-c(1:10481407,11333729:11449312), ]

#vehicle\_df <- vehicles[-c(1:13986393, 15178454:15345829), ]

View(accident)

View(casualties\_df)

View(vehicle\_df)

# Save new accident data

write.csv(accident,"C://Users//LabStudent-55-706949//Desktop//data quality/data//accident.csv", row.names = FALSE)

# Save new accident data

write.csv(casualties\_df,"C://Users//LabStudent-55-706949//Desktop//ADMP GROUP 3 Datasets//casualty.csv", row.names = FALSE)

# Save new accident data

write.csv(vehicle\_df,"C://Users//LabStudent-55-706949//Desktop//ADMP GROUP 3 Datasets//vehicles.csv", row.names = FALSE)

# Transforming selected column record from numeric to text

# Accients Dataset: Column "Light Conditions"

accident$light\_conditions[accident$light\_conditions == 1] <- "Daylight"

accident$light\_conditions[accident$light\_conditions == 4] <- "Darkness - lights lit"

accident$light\_conditions[accident$light\_conditions == 5] <- "Darkness - lights unlit"

accident$light\_conditions[accident$light\_conditions == 6] <- "Darkness - no lighting"

accident$light\_conditions[accident$light\_conditions == 7] <- "Darkness - lighting unknown"

accident$light\_conditions[accident$light\_conditions == "Data missing or out of range"] <- "NA"

View(accident)

# Column "weather\_conditions"

accident$weather\_conditions[accident$weather\_conditions == 1] <- "Fine no high winds"

accident$weather\_conditions[accident$weather\_conditions == 2] <- "Raining no high winds"

accident$weather\_conditions[accident$weather\_conditions == 3] <- "Snowing no high winds"

accident$weather\_conditions[accident$weather\_conditions == 4] <- "Fine + high winds"

accident$weather\_conditions[accident$weather\_conditions == 5] <- "Raining + high winds"

accident$weather\_conditions[accident$weather\_conditions == 6] <- "Snowing + high winds"

accident$weather\_conditions[accident$weather\_conditions == 7] <- "Fog or mist"

accident$weather\_conditions[accident$weather\_conditions == 8] <- "Other"

accident$weather\_conditions[accident$weather\_conditions == 9] <- "Unknown"

# Column "road\_surface\_condition

accident$road\_surface\_conditions[accident$road\_surface\_conditions == 1] <- "Dry"

accident$road\_surface\_conditions[accident$road\_surface\_conditions == 2] <- "Wet or damp"

accident$road\_surface\_conditions[accident$road\_surface\_conditions == 3] <- "Snow"

accident$road\_surface\_conditions[accident$road\_surface\_conditions == 4] <- "Frost or ice"

accident$road\_surface\_conditions[accident$road\_surface\_conditions == 5] <- "Flood over 3cm deep"

accident$road\_surface\_conditions[accident$road\_surface\_conditions == 6] <- "Oil or diesel"

accident$road\_surface\_conditions[accident$road\_surface\_conditions == 7] <- "Mud"

accident$road\_surface\_conditions[accident$road\_surface\_conditions == -1] <- "NA"

accident$road\_surface\_conditions[accident$road\_surface\_conditions == 9] <- "&self reported"

View(accident)

#Column "accident\_severity"

accident$accident\_severity[accident$accident\_severity == 1] <- "Fatal"

accident$accident\_severity[accident$accident\_severity == 2] <- "Serious"

accident$accident\_severity[accident$accident\_severity == 3] <- "Slight"

#Column "day of the week"

accident$day\_of\_week[accident$day\_of\_week == 1] <- "Monday"

accident$day\_of\_week[accident$day\_of\_week == 2] <- "Tuesday"

accident$day\_of\_week[accident$day\_of\_week == 3] <- "Wednesday"

accident$day\_of\_week[accident$day\_of\_week == 4] <- "Thusday"

accident$day\_of\_week[accident$day\_of\_week == 5] <- "Friday"

accident$day\_of\_week[accident$day\_of\_week == 6] <- "Satuday"

accident$day\_of\_week[accident$day\_of\_week == 7] <- "Sunday"

accident<- read.csv("C:/Users/LabStudent-55-706949/Desktop/accident.accident.csv")

#local\_authority\_district

accident$local\_authority\_district[accident$local\_authority\_district == 1]<-"Westminster"

accident$local\_authority\_district[accident$local\_authority\_district == 2]<-"Camden"

accident$local\_authority\_district[accident$local\_authority\_district == 3]<- "Islington"

accident$local\_authority\_district[accident$local\_authority\_district == 4 ] <- "Hackney"

accident $local\_authority\_district [accident$local\_authority\_district == 5 ] <- "Tower Hamlets"

accident $local\_authority\_district [accident$local\_authority\_district == 6 ] <- "Greenwich"

accident $local\_authority\_district [accident$local\_authority\_district == 7 ] <- "Lewisham"

accident $local\_authority\_district [accident$local\_authority\_district == 8 ] <- "Southwark"

accident $local\_authority\_district [accident$local\_authority\_district == 9 ] <- "Lambeth"

accident $local\_authority\_district [accident$local\_authority\_district == 10 ] <- "Wandsworth"

accident $local\_authority\_district [accident$local\_authority\_district == 11 ] <- "Hammersmith and Fulham"

accident $local\_authority\_district [accident$local\_authority\_district == 12 ] <- "Kensington and Chelsea"

accident $local\_authority\_district [accident$local\_authority\_district == 13 ] <- "Waltham Forest"

accident $local\_authority\_district [accident$local\_authority\_district == 14 ] <- "Redbridge"

accident $local\_authority\_district [accident$local\_authority\_district == 15 ] <- "Havering"

accident $local\_authority\_district [accident$local\_authority\_district == 16 ] <- "Barking and Dagenham"

accident $local\_authority\_district [accident$local\_authority\_district == 17 ] <- "Newham"

accident $local\_authority\_district [accident$local\_authority\_district == 18 ] <- "Bexley"

accident $local\_authority\_district [accident$local\_authority\_district == 19 ] <- "Bromley"

accident $local\_authority\_district [accident$local\_authority\_district == 20 ] <- "Croydon"

accident $local\_authority\_district [accident$local\_authority\_district == 21 ] <- "Sutton"

accident $local\_authority\_district [accident$local\_authority\_district == 22 ] <- "Merton"

accident $local\_authority\_district [accident$local\_authority\_district == 23 ] <- "Kingston upon Thames"

accident $local\_authority\_district [accident$local\_authority\_district == 24 ] <- "Richmond upon Thames"

accident $local\_authority\_district [accident$local\_authority\_district == 25 ] <- "Hounslow"

accident $local\_authority\_district [accident$local\_authority\_district == 26 ] <- "Hillingdon"

accident $local\_authority\_district [accident$local\_authority\_district == 27 ] <- "Ealing"

accident $local\_authority\_district [accident$local\_authority\_district == 28 ] <- "Brent"

accident $local\_authority\_district [accident$local\_authority\_district == 29 ] <- "Harrow"

accident $local\_authority\_district [accident$local\_authority\_district == 30 ] <- "Barnet"

accident $local\_authority\_district [accident$local\_authority\_district == 31 ] <- "Haringey"

accident $local\_authority\_district [accident$local\_authority\_district == 32 ] <- "Enfield"

accident $local\_authority\_district [accident$local\_authority\_district == 33 ] <- "Hertsmere"

accident $local\_authority\_district [accident$local\_authority\_district == 38 ] <- "Epsom and Ewell"

accident $local\_authority\_district [accident$local\_authority\_district == 40 ] <- "Spelthorne"

accident $local\_authority\_district [accident$local\_authority\_district == 57 ] <- "London Airport (Heathrow)"

accident $local\_authority\_district [accident$local\_authority\_district == 60 ] <- "Allerdale"

accident $local\_authority\_district [accident$local\_authority\_district == 61 ] <- "Barrow-in-Furness"

accident $local\_authority\_district [accident$local\_authority\_district == 62 ] <- "Carlisle"

accident $local\_authority\_district [accident$local\_authority\_district == 63 ] <- "Copeland"

accident $local\_authority\_district [accident$local\_authority\_district == 64 ] <- "Eden"

accident $local\_authority\_district [accident$local\_authority\_district == 65 ] <- "South Lakeland"

accident $local\_authority\_district [accident$local\_authority\_district == 70 ] <- "Blackburn with Darwen"

accident $local\_authority\_district [accident$local\_authority\_district == 71 ] <- "Blackpool"

accident $local\_authority\_district [accident$local\_authority\_district == 72 ] <- "Burnley"

accident $local\_authority\_district [accident$local\_authority\_district == 73 ] <- "Chorley"

accident $local\_authority\_district [accident$local\_authority\_district == 74 ] <- "Fylde"

accident $local\_authority\_district [accident$local\_authority\_district == 75 ] <- "Hyndburn"

accident $local\_authority\_district [accident$local\_authority\_district == 76 ] <- "Lancaster"

accident $local\_authority\_district [accident$local\_authority\_district == 77 ] <- "Pendle"

accident $local\_authority\_district [accident$local\_authority\_district == 79 ] <- "Preston"

accident $local\_authority\_district [accident$local\_authority\_district == 80 ] <- "Ribble Valley"

accident $local\_authority\_district [accident$local\_authority\_district == 82 ] <- "Rossendale"

accident $local\_authority\_district [accident$local\_authority\_district == 83 ] <- "South Ribble"

accident $local\_authority\_district [accident$local\_authority\_district == 84 ] <- "West Lancashire"

accident $local\_authority\_district [accident$local\_authority\_district == 85 ] <- "Wyre"

accident $local\_authority\_district [accident$local\_authority\_district == 90 ] <- "Knowsley"

accident $local\_authority\_district [accident$local\_authority\_district == 91 ] <- "Liverpool"

accident $local\_authority\_district [accident$local\_authority\_district == 92 ] <- "St. Helens"

accident $local\_authority\_district [accident$local\_authority\_district == 93 ] <- "Sefton"

accident $local\_authority\_district [accident$local\_authority\_district == 95 ] <- "Wirral"

accident $local\_authority\_district [accident$local\_authority\_district == 100 ] <- "Bolton"

accident $local\_authority\_district [accident$local\_authority\_district == 101 ] <- "Bury"

accident $local\_authority\_district [accident$local\_authority\_district == 102 ] <- "Manchester"

accident $local\_authority\_district [accident$local\_authority\_district == 104 ] <- "Oldham"

accident $local\_authority\_district [accident$local\_authority\_district == 106 ] <- "Rochdale"

accident $local\_authority\_district [accident$local\_authority\_district == 107 ] <- "Salford"

accident $local\_authority\_district [accident$local\_authority\_district == 109 ] <- "Stockport"

accident $local\_authority\_district [accident$local\_authority\_district == 110 ] <- "Tameside"

accident $local\_authority\_district [accident$local\_authority\_district == 112 ] <- "Trafford"

accident $local\_authority\_district [accident$local\_authority\_district == 114 ] <- "Wigan"

accident $local\_authority\_district [accident$local\_authority\_district == 120 ] <- "Chester"

accident $local\_authority\_district [accident$local\_authority\_district == 121 ] <- "Congleton"

accident $local\_authority\_district [accident$local\_authority\_district == 122 ] <- "Crewe and Nantwich"

accident $local\_authority\_district [accident$local\_authority\_district == 123 ] <- "Ellesmere Port and Neston"

accident $local\_authority\_district [accident$local\_authority\_district == 124 ] <- "Halton"

accident $local\_authority\_district [accident$local\_authority\_district == 126 ] <- "Macclesfield"

accident $local\_authority\_district [accident$local\_authority\_district == 127 ] <- "Vale Royal"

accident $local\_authority\_district [accident$local\_authority\_district == 128 ] <- "Warrington"

accident $local\_authority\_district [accident$local\_authority\_district == 129 ] <- "Cheshire East"

accident $local\_authority\_district [accident$local\_authority\_district == 130 ] <- "Cheshire West and Chester"

accident $local\_authority\_district [accident$local\_authority\_district == 139 ] <- "Northumberland"

accident $local\_authority\_district [accident$local\_authority\_district == 140 ] <- "Alnwick"

accident $local\_authority\_district [accident$local\_authority\_district == 141 ] <- "Berwick-upon-Tweed"

accident $local\_authority\_district [accident$local\_authority\_district == 142 ] <- "Blyth Valley"

accident $local\_authority\_district [accident$local\_authority\_district == 143 ] <- "Castle Morpeth"

accident $local\_authority\_district [accident$local\_authority\_district == 144 ] <- "Tynedale"

accident $local\_authority\_district [accident$local\_authority\_district == 145 ] <- "Wansbeck"

accident $local\_authority\_district [accident$local\_authority\_district == 146 ] <- "Gateshead"

accident $local\_authority\_district [accident$local\_authority\_district == 147 ] <- "Newcastle upon Tyne"

accident $local\_authority\_district [accident$local\_authority\_district == 148 ] <- "North Tyneside"

accident $local\_authority\_district [accident$local\_authority\_district == 149 ] <- "South Tyneside"

accident $local\_authority\_district [accident$local\_authority\_district == 150 ] <- "Sunderland"

accident $local\_authority\_district [accident$local\_authority\_district == 160 ] <- "Chester-le-Street"

accident $local\_authority\_district [accident$local\_authority\_district == 161 ] <- "Darlington"

accident $local\_authority\_district [accident$local\_authority\_district == 162 ] <- "Derwentside"

accident $local\_authority\_district [accident$local\_authority\_district == 163 ] <- "Durham"

accident $local\_authority\_district [accident$local\_authority\_district == 164 ] <- "Easington"

accident $local\_authority\_district [accident$local\_authority\_district == 165 ] <- "Sedgefield"

accident $local\_authority\_district [accident$local\_authority\_district == 166 ] <- "Teesdale"

accident $local\_authority\_district [accident$local\_authority\_district == 168 ] <- "Wear Valley"

accident $local\_authority\_district [accident$local\_authority\_district == 169 ] <- "County Durham"

accident $local\_authority\_district [accident$local\_authority\_district == 180 ] <- "Craven"

accident $local\_authority\_district [accident$local\_authority\_district == 181 ] <- "Hambleton"

accident $local\_authority\_district [accident$local\_authority\_district == 182 ] <- "Harrogate"

accident $local\_authority\_district [accident$local\_authority\_district == 184 ] <- "Richmondshire"

accident $local\_authority\_district [accident$local\_authority\_district == 185 ] <- "Ryedale"

accident $local\_authority\_district [accident$local\_authority\_district == 186 ] <- "Scarborough"

accident $local\_authority\_district [accident$local\_authority\_district == 187 ] <- "Selby"

accident $local\_authority\_district [accident$local\_authority\_district == 189 ] <- "York"

accident $local\_authority\_district [accident$local\_authority\_district == 200 ] <- "Bradford"

accident $local\_authority\_district [accident$local\_authority\_district == 202 ] <- "Calderdale"

accident $local\_authority\_district [accident$local\_authority\_district == 203 ] <- "Kirklees"

accident $local\_authority\_district [accident$local\_authority\_district == 204 ] <- "Leeds"

accident $local\_authority\_district [accident$local\_authority\_district == 206 ] <- "Wakefield"

accident $local\_authority\_district [accident$local\_authority\_district == 210 ] <- "Barnsley"

accident $local\_authority\_district [accident$local\_authority\_district == 211 ] <- "Doncaster"

accident $local\_authority\_district [accident$local\_authority\_district == 213 ] <- "Rotherham"

accident $local\_authority\_district [accident$local\_authority\_district == 215 ] <- "Sheffield"

accident $local\_authority\_district [accident$local\_authority\_district == 228 ] <- "Kingston upon Hull, City of"

accident $local\_authority\_district [accident$local\_authority\_district == 231 ] <- "East Riding of Yorkshire"

accident $local\_authority\_district [accident$local\_authority\_district == 232 ] <- "North Lincolnshire"

accident $local\_authority\_district [accident$local\_authority\_district == 233 ] <- "North East Lincolnshire"

accident $local\_authority\_district [accident$local\_authority\_district == 240 ] <- "Hartlepool"

accident $local\_authority\_district [accident$local\_authority\_district == 241 ] <- "Redcar and Cleveland"

accident $local\_authority\_district [accident$local\_authority\_district == 243 ] <- "Middlesbrough"

accident $local\_authority\_district [accident$local\_authority\_district == 245 ] <- "Stockton-on-Tees"

accident $local\_authority\_district [accident$local\_authority\_district == 250 ] <- "Cannock Chase"

accident $local\_authority\_district [accident$local\_authority\_district == 251 ] <- "East Staffordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 252 ] <- "Lichfield"

accident $local\_authority\_district [accident$local\_authority\_district == 253 ] <- "Newcastle-under-Lyme"

accident $local\_authority\_district [accident$local\_authority\_district == 254 ] <- "South Staffordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 255 ] <- "Stafford"

accident $local\_authority\_district [accident$local\_authority\_district == 256 ] <- "Staffordshire Moorlands"

accident $local\_authority\_district [accident$local\_authority\_district == 257 ] <- "Stoke-on-Trent"

accident $local\_authority\_district [accident$local\_authority\_district == 258 ] <- "Tamworth"

accident $local\_authority\_district [accident$local\_authority\_district == 270 ] <- "Bromsgrove"

accident $local\_authority\_district [accident$local\_authority\_district == 273 ] <- "Malvern Hills"

accident $local\_authority\_district [accident$local\_authority\_district == 274 ] <- "Redditch"

accident $local\_authority\_district [accident$local\_authority\_district == 276 ] <- "Worcester"

accident $local\_authority\_district [accident$local\_authority\_district == 277 ] <- "Wychavon"

accident $local\_authority\_district [accident$local\_authority\_district == 278 ] <- "Wyre Forest"

accident $local\_authority\_district [accident$local\_authority\_district == 279 ] <- "Bridgnorth"

accident $local\_authority\_district [accident$local\_authority\_district == 280 ] <- "North Shropshire"

accident $local\_authority\_district [accident$local\_authority\_district == 281 ] <- "Oswestry"

accident $local\_authority\_district [accident$local\_authority\_district == 282 ] <- "Shrewsbury and Atcham"

accident $local\_authority\_district [accident$local\_authority\_district == 283 ] <- "South Shropshire"

accident $local\_authority\_district [accident$local\_authority\_district == 284 ] <- "Telford and Wrekin"

accident $local\_authority\_district [accident$local\_authority\_district == 285 ] <- "Herefordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 286 ] <- "Shropshire"

accident $local\_authority\_district [accident$local\_authority\_district == 290 ] <- "North Warwickshire"

accident $local\_authority\_district [accident$local\_authority\_district == 291 ] <- "Nuneaton and Bedworth"

accident $local\_authority\_district [accident$local\_authority\_district == 292 ] <- "Rugby"

accident $local\_authority\_district [accident$local\_authority\_district == 293 ] <- "Stratford-upon-Avon"

accident $local\_authority\_district [accident$local\_authority\_district == 294 ] <- "Warwick"

accident $local\_authority\_district [accident$local\_authority\_district == 300 ] <- "Birmingham"

accident $local\_authority\_district [accident$local\_authority\_district == 302 ] <- "Coventry"

accident $local\_authority\_district [accident$local\_authority\_district == 303 ] <- "Dudley"

accident $local\_authority\_district [accident$local\_authority\_district == 305 ] <- "Sandwell"

accident $local\_authority\_district [accident$local\_authority\_district == 306 ] <- "Solihull"

accident $local\_authority\_district [accident$local\_authority\_district == 307 ] <- "Walsall"

accident $local\_authority\_district [accident$local\_authority\_district == 309 ] <- "Wolverhampton"

accident $local\_authority\_district [accident$local\_authority\_district == 320 ] <- "Amber Valley"

accident $local\_authority\_district [accident$local\_authority\_district == 321 ] <- "Bolsover"

accident $local\_authority\_district [accident$local\_authority\_district == 322 ] <- "Chesterfield"

accident $local\_authority\_district [accident$local\_authority\_district == 323 ] <- "Derby"

accident $local\_authority\_district [accident$local\_authority\_district == 324 ] <- "Erewash"

accident $local\_authority\_district [accident$local\_authority\_district == 325 ] <- "High Peak"

accident $local\_authority\_district [accident$local\_authority\_district == 327 ] <- "North East Derbyshire"

accident $local\_authority\_district [accident$local\_authority\_district == 328 ] <- "South Derbyshire"

accident $local\_authority\_district [accident$local\_authority\_district == 329 ] <- "Derbyshire Dales"

accident $local\_authority\_district [accident$local\_authority\_district == 340 ] <- "Ashfield"

accident $local\_authority\_district [accident$local\_authority\_district == 341 ] <- "Bassetlaw"

accident $local\_authority\_district [accident$local\_authority\_district == 342 ] <- "Broxtowe"

accident $local\_authority\_district [accident$local\_authority\_district == 343 ] <- "Gedling"

accident $local\_authority\_district [accident$local\_authority\_district == 344 ] <- "Mansfield"

accident $local\_authority\_district [accident$local\_authority\_district == 345 ] <- "Newark and Sherwood"

accident $local\_authority\_district [accident$local\_authority\_district == 346 ] <- "Nottingham"

accident $local\_authority\_district [accident$local\_authority\_district == 347 ] <- "Rushcliffe"

accident $local\_authority\_district [accident$local\_authority\_district == 350 ] <- "Boston"

accident $local\_authority\_district [accident$local\_authority\_district == 351 ] <- "East Lindsey"

accident $local\_authority\_district [accident$local\_authority\_district == 352 ] <- "Lincoln"

accident $local\_authority\_district [accident$local\_authority\_district == 353 ] <- "North Kesteven"

accident $local\_authority\_district [accident$local\_authority\_district == 354 ] <- "South Holland"

accident $local\_authority\_district [accident$local\_authority\_district == 355 ] <- "South Kesteven"

accident $local\_authority\_district [accident$local\_authority\_district == 356 ] <- "West Lindsey"

accident $local\_authority\_district [accident$local\_authority\_district == 360 ] <- "Blaby"

accident $local\_authority\_district [accident$local\_authority\_district == 361 ] <- "Hinckley and Bosworth"

accident $local\_authority\_district [accident$local\_authority\_district == 362 ] <- "Charnwood"

accident $local\_authority\_district [accident$local\_authority\_district == 363 ] <- "Harborough"

accident $local\_authority\_district [accident$local\_authority\_district == 364 ] <- "Leicester"

accident $local\_authority\_district [accident$local\_authority\_district == 365 ] <- "Melton"

accident $local\_authority\_district [accident$local\_authority\_district == 366 ] <- "North West Leicestershire"

accident $local\_authority\_district [accident$local\_authority\_district == 367 ] <- "Oadby and Wigston"

accident $local\_authority\_district [accident$local\_authority\_district == 368 ] <- "Rutland"

accident $local\_authority\_district [accident$local\_authority\_district == 380 ] <- "Corby"

accident $local\_authority\_district [accident$local\_authority\_district == 381 ] <- "Daventry"

accident $local\_authority\_district [accident$local\_authority\_district == 382 ] <- "East Northamptonshire"

accident $local\_authority\_district [accident$local\_authority\_district == 383 ] <- "Kettering"

accident $local\_authority\_district [accident$local\_authority\_district == 384 ] <- "Northampton"

accident $local\_authority\_district [accident$local\_authority\_district == 385 ] <- "South Northamptonshire"

accident $local\_authority\_district [accident$local\_authority\_district == 386 ] <- "Wellingborough"

accident $local\_authority\_district [accident$local\_authority\_district == 390 ] <- "Cambridge"

accident $local\_authority\_district [accident$local\_authority\_district == 391 ] <- "East Cambridgeshire"

accident $local\_authority\_district [accident$local\_authority\_district == 392 ] <- "Fenland"

accident $local\_authority\_district [accident$local\_authority\_district == 393 ] <- "Huntingdonshire"

accident $local\_authority\_district [accident$local\_authority\_district == 394 ] <- "Peterborough"

accident $local\_authority\_district [accident$local\_authority\_district == 395 ] <- "South Cambridgeshire"

accident $local\_authority\_district [accident$local\_authority\_district == 400 ] <- "Breckland"

accident $local\_authority\_district [accident$local\_authority\_district == 401 ] <- "Broadland"

accident $local\_authority\_district [accident$local\_authority\_district == 402 ] <- "Great Yarmouth"

accident $local\_authority\_district [accident$local\_authority\_district == 404 ] <- "Norwich"

accident $local\_authority\_district [accident$local\_authority\_district == 405 ] <- "North Norfolk"

accident $local\_authority\_district [accident$local\_authority\_district == 406 ] <- "South Norfolk"

accident $local\_authority\_district [accident$local\_authority\_district == 407 ] <- "King's Lynn and West Norfolk"

accident $local\_authority\_district [accident$local\_authority\_district == 410 ] <- "Babergh"

accident $local\_authority\_district [accident$local\_authority\_district == 411 ] <- "Forest Heath"

accident $local\_authority\_district [accident$local\_authority\_district == 412 ] <- "Ipswich"

accident $local\_authority\_district [accident$local\_authority\_district == 413 ] <- "Mid Suffolk"

accident $local\_authority\_district [accident$local\_authority\_district == 414 ] <- "St. Edmundsbury"

accident $local\_authority\_district [accident$local\_authority\_district == 415 ] <- "Suffolk Coastal"

accident $local\_authority\_district [accident$local\_authority\_district == 416 ] <- "Waveney"

accident $local\_authority\_district [accident$local\_authority\_district == 420 ] <- "Bedford"

accident $local\_authority\_district [accident$local\_authority\_district == 421 ] <- "Luton"

accident $local\_authority\_district [accident$local\_authority\_district == 422 ] <- "Mid Bedfordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 423 ] <- "South Bedfordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 424 ] <- "Central Bedfordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 430 ] <- "Broxbourne"

accident $local\_authority\_district [accident$local\_authority\_district == 431 ] <- "Dacorum"

accident $local\_authority\_district [accident$local\_authority\_district == 432 ] <- "East Hertfordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 433 ] <- "North Hertfordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 434 ] <- "St. Albans"

accident $local\_authority\_district [accident$local\_authority\_district == 435 ] <- "Stevenage"

accident $local\_authority\_district [accident$local\_authority\_district == 436 ] <- "Three Rivers"

accident $local\_authority\_district [accident$local\_authority\_district == 437 ] <- "Watford"

accident $local\_authority\_district [accident$local\_authority\_district == 438 ] <- "Welwyn Hatfield"

accident $local\_authority\_district [accident$local\_authority\_district == 450 ] <- "Basildon"

accident $local\_authority\_district [accident$local\_authority\_district == 451 ] <- "Braintree"

accident $local\_authority\_district [accident$local\_authority\_district == 452 ] <- "Brentwood"

accident $local\_authority\_district [accident$local\_authority\_district == 453 ] <- "Castle Point"

accident $local\_authority\_district [accident$local\_authority\_district == 454 ] <- "Chelmsford"

accident $local\_authority\_district [accident$local\_authority\_district == 455 ] <- "Colchester"

accident $local\_authority\_district [accident$local\_authority\_district == 456 ] <- "Epping Forest"

accident $local\_authority\_district [accident$local\_authority\_district == 457 ] <- "Harlow"

accident $local\_authority\_district [accident$local\_authority\_district == 458 ] <- "Maldon"

accident $local\_authority\_district [accident$local\_authority\_district == 459 ] <- "Rochford"

accident $local\_authority\_district [accident$local\_authority\_district == 460 ] <- "Southend-on-Sea"

accident $local\_authority\_district [accident$local\_authority\_district == 461 ] <- "Tendring"

accident $local\_authority\_district [accident$local\_authority\_district == 462 ] <- "Thurrock"

accident $local\_authority\_district [accident$local\_authority\_district == 463 ] <- "Uttlesford"

accident $local\_authority\_district [accident$local\_authority\_district == 470 ] <- "Bracknell Forest"

accident $local\_authority\_district [accident$local\_authority\_district == 471 ] <- "West Berkshire"

accident $local\_authority\_district [accident$local\_authority\_district == 472 ] <- "Reading"

accident $local\_authority\_district [accident$local\_authority\_district == 473] <- "Slough"

accident $local\_authority\_district [accident$local\_authority\_district == 474 ] <- "Windsor and Maidenhead"

accident $local\_authority\_district [accident$local\_authority\_district == 475 ] <- "Wokingham"

accident $local\_authority\_district [accident$local\_authority\_district == 476 ] <- "Aylesbury Vale"

accident $local\_authority\_district [accident$local\_authority\_district == 477 ] <- "South Bucks"

accident $local\_authority\_district [accident$local\_authority\_district == 478 ] <- "Chiltern"

accident $local\_authority\_district [accident$local\_authority\_district == 479 ] <- "Milton Keynes"

accident $local\_authority\_district [accident$local\_authority\_district == 480 ] <- "Wycombe"

accident $local\_authority\_district [accident$local\_authority\_district == 481 ] <- "Cherwell"

accident $local\_authority\_district [accident$local\_authority\_district == 482 ] <- "Oxford"

accident $local\_authority\_district [accident$local\_authority\_district == 483 ] <- "Vale of White Horse"

accident $local\_authority\_district [accident$local\_authority\_district == 484 ] <- "South Oxfordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 485 ] <- "West Oxfordshire"

accident $local\_authority\_district [accident$local\_authority\_district == 490 ] <- "Basingstoke and Deane"

accident $local\_authority\_district [accident$local\_authority\_district == 491 ] <- "Eastleigh"

accident $local\_authority\_district [accident$local\_authority\_district == 492 ] <- "Fareham"

accident $local\_authority\_district [accident$local\_authority\_district == 493 ] <- "Gosport"

accident $local\_authority\_district [accident$local\_authority\_district == 494 ] <- "Hart"

accident $local\_authority\_district [accident$local\_authority\_district == 495 ] <- "Havant"

accident $local\_authority\_district [accident$local\_authority\_district == 496] <- "New Forest"

accident $local\_authority\_district [accident$local\_authority\_district == 497 ] <- "East Hampshire"

accident $local\_authority\_district [accident$local\_authority\_district == 498 ] <- "Portsmouth"

accident $local\_authority\_district [accident$local\_authority\_district == 499 ] <- "Rushmoor"

accident $local\_authority\_district [accident$local\_authority\_district == 500 ] <- "Southampton"

accident $local\_authority\_district [accident$local\_authority\_district == 501] <- "Test Valley"

accident $local\_authority\_district [accident$local\_authority\_district == 502 ] <- "Winchester"

accident $local\_authority\_district [accident$local\_authority\_district == 505 ] <- "Isle of Wight"

accident $local\_authority\_district [accident$local\_authority\_district == 510 ] <- "Elmbridge"

accident $local\_authority\_district [accident$local\_authority\_district == 511 ] <- "Guildford"

accident $local\_authority\_district [accident$local\_authority\_district == 512 ] <- "Mole Valley"

accident $local\_authority\_district [accident$local\_authority\_district == 513 ] <- "Reigate and Banstead"

accident $local\_authority\_district [accident$local\_authority\_district == 514 ] <- "Runnymede"

accident $local\_authority\_district [accident$local\_authority\_district == 515 ] <- "Surrey Heath"

accident $local\_authority\_district [accident$local\_authority\_district == 516 ] <- "Tandridge"

accident $local\_authority\_district [accident$local\_authority\_district == 517 ] <- "Waverley"

accident $local\_authority\_district [accident$local\_authority\_district == 518 ] <- "Woking"

accident $local\_authority\_district [accident$local\_authority\_district == 530 ] <- "Ashford"

accident $local\_authority\_district [accident$local\_authority\_district == 531 ] <- "Canterbury"

accident $local\_authority\_district [accident$local\_authority\_district == 532 ] <- "Dartford"

accident $local\_authority\_district [accident$local\_authority\_district == 533 ] <- "Dover"

accident $local\_authority\_district [accident$local\_authority\_district == 535 ] <- "Gravesham"

accident $local\_authority\_district [accident$local\_authority\_district == 536 ] <- "Maidstone"

accident $local\_authority\_district [accident$local\_authority\_district == 538 ] <- "Sevenoaks"

accident $local\_authority\_district [accident$local\_authority\_district == 539 ] <- "Shepway"

accident $local\_authority\_district [accident$local\_authority\_district == 540 ] <- "Swale"

accident $local\_authority\_district [accident$local\_authority\_district == 541 ] <- "Thanet"

accident $local\_authority\_district [accident$local\_authority\_district == 542 ] <- "Tonbridge and Malling"

accident $local\_authority\_district [accident$local\_authority\_district == 543 ] <- "Tunbridge Wells"

accident $local\_authority\_district [accident$local\_authority\_district == 544 ] <- "Medway"

accident $local\_authority\_district [accident$local\_authority\_district == 551 ] <- "Eastbourne"

accident $local\_authority\_district [accident$local\_authority\_district == 552 ] <- "Hastings"

accident $local\_authority\_district [accident$local\_authority\_district == 554 ] <- "Lewes"

accident $local\_authority\_district [accident$local\_authority\_district == 555 ] <- "Rother"

accident $local\_authority\_district [accident$local\_authority\_district == 556 ] <- "Wealden"

accident $local\_authority\_district [accident$local\_authority\_district == 557 ] <- "Adur"

accident $local\_authority\_district [accident$local\_authority\_district == 558 ] <- "Arun"

accident $local\_authority\_district [accident$local\_authority\_district == 559 ] <- "Chichester"

accident $local\_authority\_district [accident$local\_authority\_district == 560 ] <- "Crawley"

accident $local\_authority\_district [accident$local\_authority\_district == 562 ] <- "Horsham"

accident $local\_authority\_district [accident$local\_authority\_district == 563 ] <- "Mid Sussex"

accident $local\_authority\_district [accident$local\_authority\_district == 564 ] <- "Worthing"

accident $local\_authority\_district [accident$local\_authority\_district == 565 ] <- "Brighton and Hove"

accident $local\_authority\_district [accident$local\_authority\_district == 570 ] <- "London"

accident $local\_authority\_district [accident$local\_authority\_district == 580 ] <- "East Devon"

accident $local\_authority\_district [accident$local\_authority\_district == 581 ] <- "Exeter"

accident $local\_authority\_district [accident$local\_authority\_district == 582 ] <- "North Devon"

accident $local\_authority\_district [accident$local\_authority\_district == 583 ] <- "Plymouth"

accident $local\_authority\_district [accident$local\_authority\_district == 584 ] <- "South Hams"

accident $local\_authority\_district [accident$local\_authority\_district == 585 ] <- "Teignbridge"

accident $local\_authority\_district [accident$local\_authority\_district == 586 ] <- "Mid Devon"

accident $local\_authority\_district [accident$local\_authority\_district == 587 ] <- "Torbay"

accident $local\_authority\_district [accident$local\_authority\_district == 588] <- "Torridge"

accident $local\_authority\_district [accident$local\_authority\_district == 589 ] <- "West Devon"

accident $local\_authority\_district [accident$local\_authority\_district == 590 ] <- "Caradon"

accident $local\_authority\_district [accident$local\_authority\_district == 591 ] <- "Carrick"

accident $local\_authority\_district [accident$local\_authority\_district == 592 ] <- "Kerrier"

accident $local\_authority\_district [accident$local\_authority\_district == 594 ] <- "North Cornwall"

accident $local\_authority\_district [accident$local\_authority\_district == 594 ] <- "Penwith"

accident $local\_authority\_district [accident$local\_authority\_district == 595 ] <- "Restormel"

accident $local\_authority\_district [accident$local\_authority\_district == 596 ] <- "Cornwall"

accident $local\_authority\_district [accident$local\_authority\_district == 601 ] <- "Bristol, City of"

accident $local\_authority\_district [accident$local\_authority\_district == 605 ] <- "North Somerset"

accident $local\_authority\_district [accident$local\_authority\_district == 606 ] <- "Mendip"

accident $local\_authority\_district [accident$local\_authority\_district == 607 ] <- "Sedgemoor"

accident $local\_authority\_district [accident$local\_authority\_district == 608 ] <- "Taunton Deane"

accident $local\_authority\_district [accident$local\_authority\_district == 609 ] <- "West Somerset"

accident $local\_authority\_district [accident$local\_authority\_district == 610 ] <- "South Somerset"

accident $local\_authority\_district [accident$local\_authority\_district == 611 ] <- "Bath and North East Somerset"

accident $local\_authority\_district [accident$local\_authority\_district == 612 ] <- "South Gloucestershire"

accident $local\_authority\_district [accident$local\_authority\_district == 620 ] <- "Cheltenham"

accident $local\_authority\_district [accident$local\_authority\_district == 621 ] <- "Cotswold"

accident $local\_authority\_district [accident$local\_authority\_district == 622 ] <- "Forest of Dean"

accident $local\_authority\_district [accident$local\_authority\_district == 623 ] <- "Gloucester"

accident $local\_authority\_district [accident$local\_authority\_district == 624 ] <- "Stroud"

accident $local\_authority\_district [accident$local\_authority\_district == 625 ] <- "Tewkesbury"

accident $local\_authority\_district [accident$local\_authority\_district == 630 ] <- "Kennet"

accident $local\_authority\_district [accident$local\_authority\_district == 631 ] <- "North Wiltshire"

accident $local\_authority\_district [accident$local\_authority\_district == 632 ] <- "Salisbury"

accident $local\_authority\_district [accident$local\_authority\_district == 633 ] <- "Swindon"

accident $local\_authority\_district [accident$local\_authority\_district == 634 ] <- "West Wiltshire"

accident $local\_authority\_district [accident$local\_authority\_district == 635 ] <- "Wiltshire"

accident $local\_authority\_district [accident$local\_authority\_district == 640 ] <- "Bournemouth"

accident $local\_authority\_district [accident$local\_authority\_district == 641 ] <- "Christchurch"

accident $local\_authority\_district [accident$local\_authority\_district == 642 ] <- "North Dorset"

accident $local\_authority\_district [accident$local\_authority\_district == 643 ] <- "Poole"

accident $local\_authority\_district [accident$local\_authority\_district == 644 ] <- "Purbeck"

accident $local\_authority\_district [accident$local\_authority\_district == 645 ] <- "West Dorset"

accident $local\_authority\_district [accident$local\_authority\_district == 646] <- "Weymouth and Portland"

accident $local\_authority\_district [accident$local\_authority\_district == 647 ] <- "East Dorset"

accident $local\_authority\_district [accident$local\_authority\_district == 720 ] <- "Isle of Anglesey"

accident $local\_authority\_district [accident$local\_authority\_district == 721 ] <- "Conwy"

accident $local\_authority\_district [accident$local\_authority\_district == 722 ] <- "Gwynedd"

accident $local\_authority\_district [accident$local\_authority\_district == 723 ] <- "Denbighshire"

accident $local\_authority\_district [accident$local\_authority\_district == 724 ] <- "Flintshire"

accident $local\_authority\_district [accident$local\_authority\_district == 725 ] <- "Wrexham"

accident $local\_authority\_district [accident$local\_authority\_district == 730 ] <- "Blaenau Gwent"

accident $local\_authority\_district [accident$local\_authority\_district == 731 ] <- "Caerphilly"

accident $local\_authority\_district [accident$local\_authority\_district == 732 ] <- "Monmouthshire"

accident $local\_authority\_district [accident$local\_authority\_district == 733 ] <- "Newport"

accident $local\_authority\_district [accident$local\_authority\_district == 734 ] <- "Torfaen"

accident $local\_authority\_district [accident$local\_authority\_district == 740 ] <- "Bridgend"

accident $local\_authority\_district [accident$local\_authority\_district == 741 ] <- "Cardiff"

accident $local\_authority\_district [accident$local\_authority\_district == 742 ] <- "Merthyr Tydfil"

accident $local\_authority\_district [accident$local\_authority\_district == 743 ] <- "Neath Port Talbot"

accident $local\_authority\_district [accident$local\_authority\_district == 744 ] <- "Rhondda, Cynon, Taff"

accident $local\_authority\_district [accident$local\_authority\_district == 745 ] <- "Swansea"

accident $local\_authority\_district [accident$local\_authority\_district == 746 ] <- "The Vale of Glamorgan"

accident $local\_authority\_district [accident$local\_authority\_district == 750 ] <- "Ceredigion"

accident $local\_authority\_district [accident$local\_authority\_district == 751 ] <- "Carmarthenshire"

accident $local\_authority\_district [accident$local\_authority\_district == 752 ] <- "Pembrokeshire"

accident $local\_authority\_district [accident$local\_authority\_district == 753 ] <- "Powys"

accident $local\_authority\_district [accident$local\_authority\_district == 910 ] <- "Aberdeen City"

accident $local\_authority\_district [accident$local\_authority\_district == 911 ] <- "Aberdeenshire"

accident $local\_authority\_district [accident$local\_authority\_district == 912 ] <- "Angus"

accident $local\_authority\_district [accident$local\_authority\_district == 913 ] <- "Argyll and Bute"

accident $local\_authority\_district [accident$local\_authority\_district == 914 ] <- "Scottish Borders"

accident $local\_authority\_district [accident$local\_authority\_district == 915 ] <- "Clackmannanshire"

accident $local\_authority\_district [accident$local\_authority\_district == 916 ] <- "West Dunbartonshire"

accident $local\_authority\_district [accident$local\_authority\_district == 917 ] <- "Dumfries and Galloway"

accident $local\_authority\_district [accident$local\_authority\_district == 918 ] <- "Dundee City"

accident $local\_authority\_district [accident$local\_authority\_district == 919 ] <- "East Ayrshire"

accident $local\_authority\_district [accident$local\_authority\_district == 920 ] <- "East Dunbartonshire"

accident $local\_authority\_district [accident$local\_authority\_district == 921 ] <- "East Lothian"

accident $local\_authority\_district [accident$local\_authority\_district == 922 ] <- "East Renfrewshire"

accident $local\_authority\_district [accident$local\_authority\_district == 923 ] <- "Edinburgh, City of"

accident $local\_authority\_district [accident$local\_authority\_district == 924 ] <- "Falkirk"

accident $local\_authority\_district [accident$local\_authority\_district == 925 ] <- "Fife"

accident $local\_authority\_district [accident$local\_authority\_district == 926 ] <- "Glasgow City"

accident $local\_authority\_district [accident$local\_authority\_district == 927 ] <- "Highland"

accident $local\_authority\_district [accident$local\_authority\_district == 928 ] <- "Inverclyde"

accident $local\_authority\_district [accident$local\_authority\_district == 929 ] <- "Midlothian"

accident $local\_authority\_district [accident$local\_authority\_district == 930 ] <- "Moray"

accident $local\_authority\_district [accident$local\_authority\_district == 931 ] <- "North Ayrshire"

accident $local\_authority\_district [accident$local\_authority\_district == 932 ] <- "North Lanarkshire"

accident $local\_authority\_district [accident$local\_authority\_district == 933 ] <- "Orkney Islands"

accident $local\_authority\_district [accident$local\_authority\_district == 934 ] <- "Perth and Kinross"

accident $local\_authority\_district [accident$local\_authority\_district == 935 ] <- "Renfrewshire"

accident $local\_authority\_district [accident$local\_authority\_district == 936 ] <- "Shetland Islands"

accident $local\_authority\_district [accident$local\_authority\_district == 937 ] <- "South Ayrshire"

accident $local\_authority\_district [accident$local\_authority\_district == 938 ] <- "South Lanarkshire"

accident $local\_authority\_district [accident$local\_authority\_district == 939 ] <- "Stirling"

accident $local\_authority\_district [accident$local\_authority\_district == 940 ] <- "West Lothian"

accident $local\_authority\_district [accident$local\_authority\_district == 941 ] <- "Western Isles"

#accident: road type

accident $road\_type [accident$road\_type == 1 ] <- "Roundabout"

accident $road\_type [accident$road\_type == 2 ] <- "one way street"

accident $road\_type [accident$road\_type == 3 ] <- "Dual carriageway"

accident $road\_type [accident$road\_type == 6 ] <- "single carriageway"

accident $road\_type [accident$road\_type == 7 ] <- "slip road"

accident $road\_type [accident$road\_type == 9 ] <- "One way street/Slip road"

#accident: junction details

accident $junction\_detail [accident$junction\_detail == 0 ] <- "Not at junction"

accident $junction\_detail [accident$junction\_detail == 1 ] <- "Roundabout"

accident $junction\_detail [accident$junction\_detail == 2 ] <- "Mini roundabout"

accident $junction\_detail [accident$junction\_detail == 3 ] <- "Tor staggered junction"

accident $junction\_detail [accident$junction\_detail == 5 ] <- "Slip road"

accident $junction\_detail [accident$junction\_detail == 6 ] <- "Crossroads"

accident $junction\_detail [accident$junction\_detail == 7 ] <- "more than 4 arms not roundabout"

accident $junction\_detail [accident$junction\_detail == 8 ] <- "Private drive or entrance"

accident $junction\_detail [accident$junction\_detail == 9 ] <- "Other junction"

accident $junction\_detail [accident$junction\_detail == 99] <- "Unknown self reported"

accident $junction\_detail [accident$junction\_detail == -1 ] <- "NA"

#accident: did police attend to the scene of the accident

accident $did\_police\_officer\_attend\_scene\_of\_accident [accident$did\_police\_officer\_attend\_scene\_of\_accident == 1 ] <- "Yes"

accident $did\_police\_officer\_attend\_scene\_of\_accident [accident$did\_police\_officer\_attend\_scene\_of\_accident == 2 ] <- "No"

accident $did\_police\_officer\_attend\_scene\_of\_accident [accident$did\_police\_officer\_attend\_scene\_of\_accident == 3 ] <- "No - accident was reported using a self completion form"

accident $did\_police\_officer\_attend\_scene\_of\_accident [accident$did\_police\_officer\_attend\_scene\_of\_accident == -1 ] <- "NA"

View(accident)

#droping unwated colums

accident <- subset (accident, select = -location\_easting\_osgr)

accident <- subset (accident, select = -c(location\_northing\_osgr,longitude,latitude,first\_road\_class,first\_road\_number,junction\_control,second\_road\_class,second\_road\_number,pedestrian\_crossing\_human\_control,pedestrian\_crossing\_physical\_facilities,special\_conditions\_at\_site,carriageway\_hazards,urban\_or\_rural\_area,trunk\_road\_flag))

View(newaccident)

newaccident <- subset (newaccident, select = -c(police\_force,did\_police\_officer\_attend\_scene\_of\_accident))

#SAVING CLEANED ACCIDENT DATASET

write.csv(accident,"C://Users//LabStudent-55-706949//Desktop//data quality/data//newaccident.csv", row.names = FALSE)

#Casualties dataset: Column "casualty\_class"

casualties$casualty\_class[casualties$casualty\_class == 1] <- "Driver or rider"

casualties$casualty\_class[casualties$casualty\_class == 2] <- "Passager"

casualties$casualty\_class[casualties$casualty\_class == 3] <- "Pedestrian"

View(casualties)

#Casualties dataset: Column "sex\_of\_casualty"

casualties$sex\_of\_casualty[casualties$sex\_of\_casualty == 1] <- "Male"

casualties$sex\_of\_casualty[casualties$sex\_of\_casualty == 2] <- "Female"

casualties$sex\_of\_casualty[casualties$sex\_of\_casualty == 9] <- "Self reported"

casualties$sex\_of\_casualty[casualties$sex\_of\_casualty == -1] <- "NA"

#casualties dataset: column"age\_of\_casualty

casualties$age\_of\_casualty[casualties$age\_of\_casualty == -1] <- "NA"

#Casualties dataset: Column "age\_band\_of\_casualty"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 1] <- "0 -5"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 2] <- "6 - 10"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 3] <- "11 - 15"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 4] <- "16 - 20"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 5] <- "21 - 25"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 6] <- "26 - 35"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 7] <- "36 - 45"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 8] <- "46 - 55"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 9] <- "56 - 65"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 10] <- "66 - 75"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == 11] <- "Over 75"

casualties$age\_band\_of\_casualty[casualties$age\_band\_of\_casualty == -1] <- "NA"

#Casualties dataset: Column "casualty\_severity"

casualties$casualty\_severity[casualties$casualty\_severity == 1] <- "Fatal"

casualties$casualty\_severity[casualties$casualty\_severity == 2] <- "Serious"

casualties$casualty\_severity[casualties$casualty\_severity == 3] <- "Slight"

#Casualties dataset: Column "casualty type"

casualties$casualty\_type[casualties$casualty\_type == 0] <- "Pedestrian"

casualties$casualty\_type[casualties$casualty\_type == 1] <- "Cyclist"

casualties$casualty\_type[casualties$casualty\_type == 2] <- "Motorcycle 50cc and under rider or passenger"

casualties$casualty\_type[casualties$casualty\_type == 3] <- "Motorcycle 125cc and under rider or passenger"

casualties$casualty\_type[casualties$casualty\_type == 4] <- "Motorcycle over 125cc and up to 500cc rider or passenger"

casualties$casualty\_type[casualties$casualty\_type == 5] <- "Motorcycle over 500cc rider or passenger"

casualties$casualty\_type[casualties$casualty\_type == 8] <- "Taxi/Private hire car occupant"

casualties$casualty\_type[casualties$casualty\_type == 9] <- "Car occupant"

casualties$casualty\_type[casualties$casualty\_type == 10] <- "Minibus (8 - 16 passenger seats) occupant"

casualties$casualty\_type[casualties$casualty\_type == 11] <- "Bus or coach occupant (17 or more pass seats)"

casualties$casualty\_type[casualties$casualty\_type == 16] <- "Horse rider"

casualties$casualty\_type[casualties$casualty\_type == 17] <- "Agricultural vehicle occupant"

casualties$casualty\_type[casualties$casualty\_type == 18] <- "Tram occupant"

casualties$casualty\_type[casualties$casualty\_type == 19] <- "Van / Goods vehicle (3.5 tonnes mgw or under) occupant"

casualties$casualty\_type[casualties$casualty\_type == 20] <- "Goods vehicle (over 3.5t. and under 7.5t.) occupant"

casualties$casualty\_type[casualties$casualty\_type == 21] <- "Goods vehicle (7.5 tonnes mgw and over) occupant"

casualties$casualty\_type[casualties$casualty\_type == 22] <- "Mobility scooter rider"

casualties$casualty\_type[casualties$casualty\_type == 23] <- "Electric motorcycle rider or passenger"

casualties$casualty\_type[casualties$casualty\_type == 90] <- "Other vehicle occupantr"

casualties$casualty\_type[casualties$casualty\_type == 97] <- "Motorcycle - unknown cc rider or passenger"

casualties$casualty\_type[casualties$casualty\_type == 98] <- "Goods vehicle (unknown weight) occupant"

casualties$casualty\_type[casualties$casualty\_type == 99] <- "Self reported"

casualties$casualty\_type[casualties$casualty\_type == 103] <- "Motorcycle - Scooter (1979-1998)"

casualties$casualty\_type[casualties$casualty\_type == 104] <- "Motorcycle (1979-1998)"

casualties$casualty\_type[casualties$casualty\_type == 105] <- "Motorcycle - Combination (1979-1998)"

casualties$casualty\_type[casualties$casualty\_type == 106] <- "Motorcycle over 125cc (1999-2004)"

casualties$casualty\_type[casualties$casualty\_type == 108] <- "Taxi (excluding private hire cars) (1979-2004)"

casualties$casualty\_type[casualties$casualty\_type == 109] <- "Car (including private hire cars) (1979-2004)"

casualties$casualty\_type[casualties$casualty\_type == 110] <- "Minibus/Motor caravan (1979-1998)"

casualties$casualty\_type[casualties$casualty\_type == 113] <- "Goods over 3.5 tonnes (1979-1998)"

#droping unwated colums

casualties <- subset (casualties, select = -c(pedestrian\_location,pedestrian\_movement,car\_passenger,bus\_or\_coach\_passenger,pedestrian\_road\_maintenance\_worker,casualty\_home\_area\_type,casualty\_imd\_decile))

View(casualties)

#SAVING CLEANED CASUALTIES DATASET

write.csv(casualties,"C://Users//LabStudent-55-706949//Desktop//data quality/data//newcasualties.csv", row.names = FALSE)

#Transforming vehicle dataset

View(vehicles)

# Vehicles Dataset: Column "vehicles type"

vehicles$vehicle\_type[vehicles$vehicle\_type == 1] <- "Pedal cycle"

vehicles$vehicle\_type[vehicles$vehicle\_type == 2] <- "Motorcycle 50cc and under"

vehicles$vehicle\_type[vehicles$vehicle\_type == 3] <- "Motorcycle 125cc and under "

vehicles$vehicle\_type[vehicles$vehicle\_type == 4] <- "Motorcycle over 125cc and up to 500cc"

vehicles$vehicle\_type[vehicles$vehicle\_type == 5] <- "Motorcycle over 500cc"

vehicles$vehicle\_type[vehicles$vehicle\_type == 8] <- "Taxi/Private hire car"

vehicles$vehicle\_type[vehicles$vehicle\_type == 9] <- "Car"

vehicles$vehicle\_type[vehicles$vehicle\_type == 10] <- "Minibus (8 - 16 passenger seats)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 11] <- "Bus or coach (17 or more pass seats)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 16] <- "Ridden horse"

vehicles$vehicle\_type[vehicles$vehicle\_type == 17] <- "Agricultural vehicle"

vehicles$vehicle\_type[vehicles$vehicle\_type == 18] <- "Tram occupant"

vehicles$vehicle\_type[vehicles$vehicle\_type == 19] <- "Van / Goods vehicle (3.5 tonnes mgw or under)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 20] <- "Goods vehicle (over 3.5t. and under 7.5t.)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 21] <- "Goods vehicle (7.5 tonnes mgw and over)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 22] <- "Mobility scooter"

vehicles$vehicle\_type[vehicles$vehicle\_type == 23] <- "Electric motorcycle "

vehicles$vehicle\_type[vehicles$vehicle\_type == 90] <- "Other vehicle "

vehicles$vehicle\_type[vehicles$vehicle\_type == 97] <- "Motorcycle - unknown cc "

vehicles$vehicle\_type[vehicles$vehicle\_type == 98] <- "Goods vehicle (unknown weight) "

vehicles$vehicle\_type[vehicles$vehicle\_type == 99] <- "Self reported"

vehicles$vehicle\_type[vehicles$vehicle\_type == 103] <- "Motorcycle - Scooter (1979-1998)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 104] <- "Motorcycle (1979-1998)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 105] <- "Motorcycle - Combination (1979-1998)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 106] <- "Motorcycle over 125cc (1999-2004)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 108] <- "Taxi (excluding private hire cars) (1979-2004)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 109] <- "Car (including private hire cars) (1979-2004)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 110] <- "Minibus/Motor caravan (1979-1998)"

vehicles$vehicle\_type[vehicles$vehicle\_type == 113] <- "Goods over 3.5 tonnes (1979-1998)"

vehicles$vehicle\_type[vehicles$vehicle\_type == -1] <- "NA"

# Vehicles Dataset: Column "Junction location"

vehicles$junction\_location[vehicles$junction\_location == 0] <- "Not a junction"

vehicles$junction\_location[vehicles$junction\_location == 1] <- "Approaching junction"

vehicles$junction\_location[vehicles$junction\_location == 2] <- "Cleared junction"

vehicles$junction\_location[vehicles$junction\_location == 3] <- "Leaving roundabout"

vehicles$junction\_location[vehicles$junction\_location == 4] <- "Entering roundabout"

vehicles$junction\_location[vehicles$junction\_location == 5] <- "Leaving main road"

vehicles$junction\_location[vehicles$junction\_location == 6] <- "Entering main road"

vehicles$junction\_location[vehicles$junction\_location == 7] <- "Entering from slip road"

vehicles$junction\_location[vehicles$junction\_location == 8] <- "Mid Junction"

vehicles$junction\_location[vehicles$junction\_location == 9] <- "self reported"

vehicles$junction\_location[vehicles$junction\_location == -1] <- "NA"

#Vehicle dataset column: journey\_purpose\_of\_driver

vehicles$journey\_purpose\_of\_driver[vehicles$journey\_purpose\_of\_driver == 1] <- "Journey as part of work"

vehicles$journey\_purpose\_of\_driver[vehicles$journey\_purpose\_of\_driver == 2] <- "Commuting to/from work"

vehicles$journey\_purpose\_of\_driver[vehicles$journey\_purpose\_of\_driver == 3] <- "taking upil to/from school"

vehicles$journey\_purpose\_of\_driver[vehicles$journey\_purpose\_of\_driver == 4] <- "Pupil riding to/from school"

vehicles$journey\_purpose\_of\_driver[vehicles$journey\_purpose\_of\_driver == 5] <- "Other"

vehicles$journey\_purpose\_of\_driver[vehicles$journey\_purpose\_of\_driver == 6] <- "Not known"

vehicles$journey\_purpose\_of\_driver[vehicles$journey\_purpose\_of\_driver == 15] <- "Other"

vehicles$journey\_purpose\_of\_driver[vehicles$journey\_purpose\_of\_driver == -1] <- "NA"

#Vehicle dataset column: sex\_of\_driver

vehicles$sex\_of\_driver[vehicles$sex\_of\_driver == 1] <- "Male"

vehicles$sex\_of\_driver[vehicles$sex\_of\_driver == 2] <- "Female"

vehicles$sex\_of\_driver[vehicles$sex\_of\_driver == 3] <- "other"

vehicles$sex\_of\_driver[vehicles$sex\_of\_driver == -1] <- "NA"

#Vehicle dataset column: age\_of\_driver

vehicles$age\_of\_driver[vehicles$age\_of\_driver == -1] <- "NA"

#Vehicle dataset column:age\_band\_of\_driver

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 1] <- "0-5"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 2] <- "6-10"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 3] <- "11-15"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 4] <- "16-20"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 5] <- "21-25"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 6] <- "26-35"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 7] <- "36-45"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 8] <- "46-55"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 9] <- "56-65"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 10] <- "66-75"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == 11] <- "over 75"

vehicles$age\_band\_of\_driver[vehicles$age\_band\_of\_driver == -1] <- "NA"

#Vehicle dataset column:propulsion\_code

vehicles$propulsion\_code[vehicles$propulsion\_code == 1] <- "petrol"

vehicles$propulsion\_code[vehicles$propulsion\_code == 2] <- "Heavy oil"

vehicles$propulsion\_code[vehicles$propulsion\_code == 3] <- "Electric"

vehicles$propulsion\_code[vehicles$propulsion\_code == 4] <- "Steam"

vehicles$propulsion\_code[vehicles$propulsion\_code == 5] <- "Gas"

vehicles$propulsion\_code[vehicles$propulsion\_code == 6] <- "Petrol/Gas (LPG)"

vehicles$propulsion\_code[vehicles$propulsion\_code == 7] <- "Gas/Bi-fuel"

vehicles$propulsion\_code[vehicles$propulsion\_code == 8] <- "Hybrid electric"

vehicles$propulsion\_code[vehicles$propulsion\_code == 9] <- "Gas Diesel"

vehicles$propulsion\_code[vehicles$propulsion\_code == 10] <- "New fuel technology"

vehicles$propulsion\_code[vehicles$propulsion\_code == 11] <- "Fuel cells"

vehicles$propulsion\_code[vehicles$propulsion\_code == 12] <- "Electric diesel"

vehicles$propulsion\_code[vehicles$propulsion\_code == -1] <- "NA"

#Vehicle dataset column:engine\_capacity\_cc

vehicles$engine\_capacity\_cc[vehicles$engine\_capacity\_cc == -1] <- "NA"

#Vehicle dataset column:age\_of\_vehicle

vehicles$age\_of\_vehicle[vehicles$age\_of\_vehicle == -1] <- "NA"

#droping unwated colums

vehicles <- subset (vehicles, select = -c(driver\_home\_area\_type,driver\_imd\_decile,generic\_make\_model,vehicle\_left\_hand\_drive,first\_point\_of\_impact,hit\_object\_off\_carriageway,vehicle\_leaving\_carriageway,hit\_object\_in\_carriageway,skidding\_and\_overturning,vehicle\_location\_restricted\_lane,vehicle\_direction\_to,vehicle\_direction\_from,vehicle\_manoeuvre,towing\_and\_articulation))

View(vehicles)

#SAVING CLEANED vehicle DATASET

write.csv(vehicles,"C://Users//LabStudent-55-706949//Desktop//data quality/data//newvehicles.csv", row.names = FALSE)

#WEATHER DATASET

yorkshireweather <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/yorkshireweather.csv")

sheffieldweather <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/sheffieldweather.csv")

View(yorkshireweather)

View(sheffieldweather)

#droping unwated colums

yorkshireweather <- subset (yorkshireweather, select = -c(tempmax,tempmin,temp,feelslikemax,feelslikemin,feelslike,dew,humidity,precip,precipprob,precipcover,snow,snowdepth,windgust,windspeed,winddir,sealevelpressure,cloudcover,visibility,solarradiation,solarenergy,uvindex,severerisk,sunrise,sunset,moonphase,stations))

View(yorkshireweather)

sheffieldweather <- subset (sheffieldweather, select = -c(tempmax,tempmin,temp,feelslikemax,feelslikemin,feelslike,dew,humidity,precip,precipprob,precipcover,snow,snowdepth,windgust,windspeed,winddir,sealevelpressure,cloudcover,visibility,solarradiation,solarenergy,uvindex,severerisk,sunrise,sunset,moonphase,stations))

View(sheffieldweather)

#SAVING CLEANED weather DATASET

write.csv(yorkshireweather,"C://Users//LabStudent-55-706949//Desktop//data quality/data//newyorkshireweather.csv", row.names = FALSE)

write.csv(sheffieldweather,"C://Users//LabStudent-55-706949//Desktop//data quality/data//newsheffieldweather.csv", row.names = FALSE)

head(newaccident)

View(casualties)

newaccident <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/newaccident.csv")

newcasualties <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/newcasualties.csv")

newsheffieldweather <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/newsheffieldweather.csv")

newvehicles <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/newvehicles.csv")

newyorkshireweather <- read.csv("C:/Users/LabStudent-55-706949/Desktop/data quality/data/newyorkshireweather.csv")

View(newaccident)

View(newcasualties)

View(newsheffieldweather)

View(newvehicles)

View(newyorkshireweather)

#DATA QUALITY CHECK

#step1

#checking the values of character variables

# Produce a Simple Frequency Count

count(newaccident, 'number\_of\_vehicles',"number\_of\_casualties","accident\_severity","day\_of\_week","light\_conditions","weather\_conditions","road\_surface\_conditions")

count(newcasualties, 'casualty\_class',"age\_of\_casualty","sex\_of\_casualty","age\_band\_of\_casualty","casualty\_severity","casualty\_type")

count(newsheffieldweather, "preciptype","conditions","desciption","icon")

count(newvehicles, "vehicle\_type","junction\_location","journey\_purpose\_of\_driver","sex\_of\_driver","age\_of\_driver","age\_band\_of\_driver","propulsion\_code","engine\_capacity\_cc","age\_of\_vehicle")

count(newyorkshireweather, "preciptype","conditions","desciption","icon")

#Horizontal Bar plots

# Read the missing cells into the counts object along with any missing values and Horizontal Bar Plots

counts <- table(newaccident$accident\_severity, useNA = "ifany")

names(counts)[is.na(names(counts))]<-"NA"

#pie(counts, main = "Accident Severity distribution")

barplot(counts, main = "accident\_severity", xlab = 'Counts', ylab = 'number\_of\_vehicles', horiz = FALSE)

#weather

counts <- table(newsheffieldweather$conditions, useNA = "ifany")

names(counts)[is.na(names(counts))]<-"NA"

#pie(counts, main = "Weather distribution")

barplot(counts, main = "conditions", xlab = 'Counts', ylab = 'Weather Condition', horiz = FALSE)

#step2

# Identify and Locate Invalid values

#Checking missing values in the 3 datasets

# ACCIDENT DATASET: Store indexes of missing values in an integer-valued vector

MissingValues = which(is.na(newaccident), arr.ind = TRUE)

# Get rownames of missing values and store in object MissVals

x = rownames(newaccident)[MissingValues[,1]]

# Get column names of missing values and store in object ColMissVals

y = colnames(newaccident)[MissingValues[,2]]

# Merge objects MissVal and ColMissVals with equal dimensions

LocatedMissingValues = paste(x, y, sep=" ")

LocatedMissingValues

# CASUALTIES DATASET: Store indexes of missing values in an integer-valued vector

CasMissVal = which(is.na(newcasualties), arr.ind = TRUE)

# Get rownames of missing values and store in object MissVals

CasMissVals = rownames(newcasualties)[CasMissVal[,1]]

# Get column names of missing values and store in object ColMissVals

CasColMissVals = colnames(newcasualties)[CasMissVal[,2]]

# Merge objects MissVal and ColMissVals with equal dimensions

CasLocatedMissesVals = paste(CasMissVals, CasColMissVals, sep = " ")

CasLocatedMissesVals

# WEATHER DATASET: Store indexes of missing values in an integer-valued vector

WeaMissVal = which(is.na(newyorkshireweather), arr.ind = TRUE)

# Get rownames of missing values and store in object MissVals

WeaMissVals = rownames(newyorkshireweather)[WeaMissVal[,1]]

# Get column names of missing values and store in object ColMissVals

WeaColMissVals = colnames(newyorkshireweather)[WeaMissVal[,2]]

# Merge objects WeaMissVal and WeaColMissVals with equal dimensions

WeaLocatedMissesVals = paste(WeaMissVals, WeaColMissVals, sep = "")

WeaLocatedMissesVals

# VEHICLE DATASET: Store indexes of missing values in an integer-valued vector

VehMissVal = which(is.na(newvehicles), arr.ind = TRUE)

# Get rownames of missing values and store in object MissVals

VehMissVals = rownames(newvehicles)[VehMissVal[,1]]

# Get column names of missing values and store in object ColMissVals

VehColMissVals = colnames(newvehicles)[VehMissVal[,2]]

# Merge objects VehMissVal and VehColMissVals with equal dimensions

VehLocatedMissesVals = paste(VehMissVals, VehColMissVals, sep = " ")

WeaLocatedMissesVals

# Check if the interested columns contains any non-numeric values

AccNonNumeric <- unlist(lapply(newaccident, is.numeric))

AccNonNumeric

newaccident[ , AccNonNumeric]

CasNonNumeric <- unlist(lapply(newcasualties, is.numeric))

CasNonNumeric

newcasualties[ , CasNonNumeric]

WeaNonNumeric <- unlist(lapply(newyorkshireweather, is.numeric))

WeaNonNumeric

newyorkshireweather[ , WeaNonNumeric]

VehNonNumeric <- unlist(lapply(vehicles, is.numeric))

VehNonNumeric

newvehicles[ , VehNonNumeric]

# List all values in non-numeric columns

newaccident[, AccNonNumeric]

newcasualties[, CasNonNumeric]

newyorkshireweather[, WeaNonNumeric]

newvehicles[, VehNonNumeric]

# Check for Non Character Values in the datasets

AccNonChar <- unlist(lapply(newaccident, is.character))

AccNonChar

newaccident[, AccNonChar]

CasNonChar <- unlist(lapply(newcasualties, is.character))

CasNonChar

newcasualties[, CasNonChar]

WeaNonChar <- unlist(lapply(newyorkshireweather, is.character))

WeaNonChar

newyorkshireweather[, WeaNonChar]

VehNonChar <- unlist(lapply(newvehicles, is.character))

VehNonChar

newvehicles[, VehNonChar]

is.na(newaccident)

which(is.na(newaccident))

#checking the values of numeric variables

#summary information for all dataset (numeric variables)

summary(newaccident)

summary(newcasualties)

summary(newvehicles)

summary(newsheffieldweather)

summary(newyorkshireweather)

#histogram for numeric variables

#accident dataset

hist(newaccident$number\_of\_casualties,col="red", ylab ="Number of casualties involved in an accident", xlab = "Count" )

hist(newaccident$number\_of\_vehicles,col="yellow",ylab ="Number of vehicles involved in an accident", xlab = "Count")

hist(newaccident$speed\_limit,col="red")

#casulties dataset

hist(newcasualties$age\_of\_casualty,col="red")

#vehicle dataset

hist(newvehicles$age\_of\_driver,col="red")

hist(newvehicles$age\_of\_vehicle,col="red")

#Identifying and locating invalid values (numeric variables)

#Accident dataset

install.package("dplyr")

newaccident %>%filter(is.na(number\_of\_casualties))

newaccident %>% filter (is.na(number\_of\_vehicles))

newaccident %>% filter (is.na(speed\_limit))

#Idenifying outliers

outliers1 <- subset(newaccident, number\_of\_casualties < 0 | number\_of\_casualties > 10000)

outliers1

outliers2 <- subset(newaccident, number\_of\_vehicles < 0 | number\_of\_vehicles > 10000)

outliers2

#Working with dates

#checking that date variable is a date

class(newaccident$date)

class(newsheffieldweather$date)

#formating the date variable as date

newaccident$date = dmy(newaccident$date)

format(newsheffieldweather$date, "%d/%m/%y")

#checking for missing values for Date

newaccident %>% filter (is.na(date))

newsheffieldweather %>% filter (is.na(datetime))

#Dealing with duplicates observation

#Identifying duplicate values of accident index (Accident dataset)

duplicate <- data.frame (table(newaccident$accident\_index))

duplicate [duplicate$Freq > 1,]

newaccident[newaccident$accident\_index %in% duplicate$var1[duplicate$Freq > 1],]

#Identifying duplicate values of accident index (casualty dataset)

duplicate1 <- data.frame (table(newcasualties$accident\_index))

duplicate1 [duplicate1$Freq > 1,]

newcasualties [newcasualties $accident\_index %in% duplicate1$var1[duplicate1$Freq > 1],]

#Identifying duplicate values of accident index (casualty dataset)

duplicate2 <- data.frame (table(newvehicles$accident\_index))

duplicate2 [duplicate2$Freq > 1,]

newvehicles [newvehicles $accident\_index %in% duplicate2$var1[duplicate2$Freq > 1],]

#remame a column

View(newaccident)

names(newaccident)[1] <-'accident\_index'

names(newcasualties)[1] <-'accident\_index'

names(newvehicles)[1] <-'accident\_index'

names(newsheffieldweather)[2] <-'date'

names(newyorkshireweather)[2] <-'date'

#Intergrity check between datasets

library("dplyr")

#identify users in accident dataset but not in casualty and vehicle dataset by accident index

anti\_join(newaccident, newcasualties, newvehicles, by = "accident\_index")

SOURCE CODE FOR BUILDING TABLES IN HIVE

Fact table with foreighkey

select dimaccident\_locationid, dimweather\_conditionid,sum(number\_of\_casualties)

from dimaccident\_location,dimweather\_condition,accidenttable

GROUP BY dimaccident\_locationid, dimweather\_conditionid;

fact tables

select accident\_year, accident\_month,accident\_location,weather\_conditions,road\_surface\_conditions,light\_conditions, sum(number\_of\_casualties)

from accidenttable

GROUP BY accident\_year, accident\_month,accident\_location,weather\_conditions,road\_surface\_conditions,light\_conditions

ORDER BY accident\_location DESC;

Dim accident severity

select row\_number() over() as dimaccident\_severityid, accident\_severity as dimaccident\_severity from accidenttable group by accident\_severity;

create table

create table dimroad\_type as(select row\_number() over() as dimroad\_typeid, road\_type as dimroad\_type from accidenttable group by road\_type);

SELECT Concat(accident\_year, accident\_monthnum) AS timeid,

accident\_year AS year,

accident\_monthnum AS month\_Numb,

accident\_month AS month\_name,

day\_of\_week

from accidenttable;

select acc\_id AS dimdriverID, accident\_severity,age\_of\_driver, sex\_of\_driver, age\_of\_vehicle, number\_of\_casualties, number\_of\_vehicle

from accidenttable;

select acc\_id AS accident\_severityID , accident\_severity

from accidenttable;

SELECT DISTINCT(accident\_severity, number\_of\_casualties)

from accidenttable;

select acc\_id AS dimdriverID, accident\_severity,age\_of\_driver, sex\_of\_driver, age\_of\_vehicle, number\_of\_casualties, number\_of\_vehicle

from accidenttable;

select row\_number() over() as raodtype\_id, road\_type as roadType

from accidenttable

group by road\_type;

Separate date into year , month, day

SELECT Substr(accident\_date, 1, 4) AS accident\_year,

Substr(accident\_date, 6, 2) AS accident\_month,

Substr(accident\_date, 8, 2) AS accident\_day,

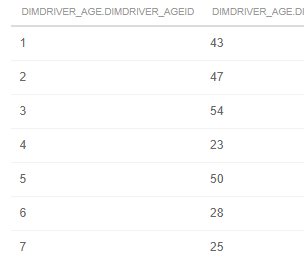
\*

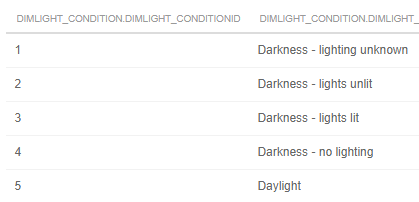
FROM tblaccident;

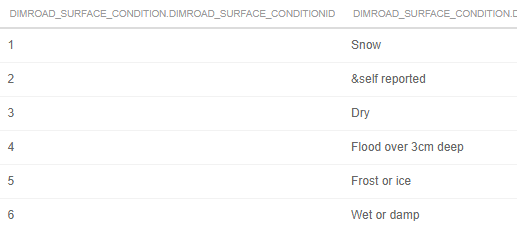
Separate the data

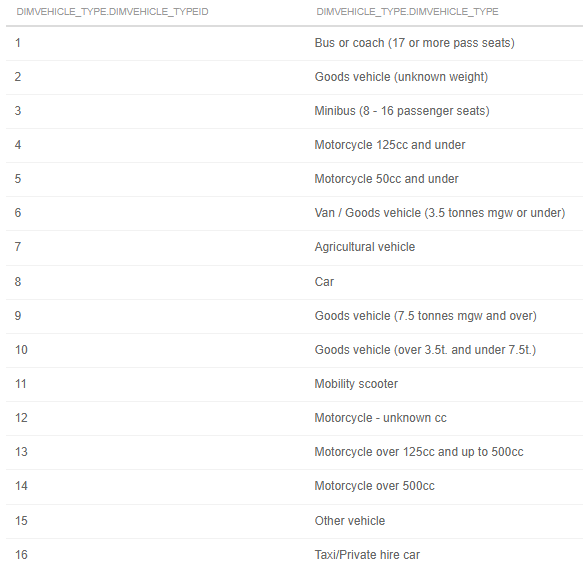
create table tblaccident as (SELECT Concat(accident\_year, "-", accident\_monthnum, "-", accident\_daynum) AS

Screenshot of FACT AND DIM tables

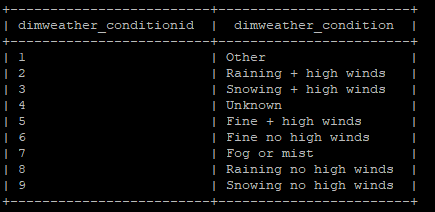
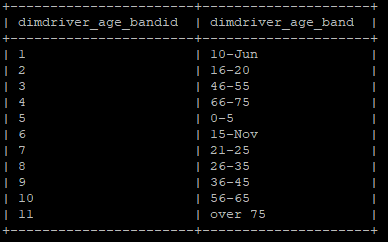
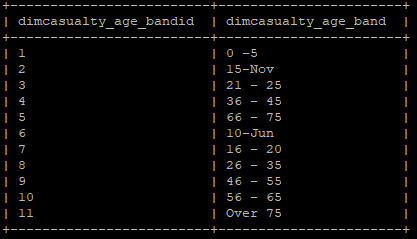
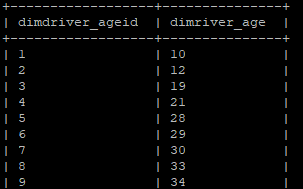
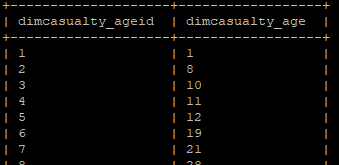
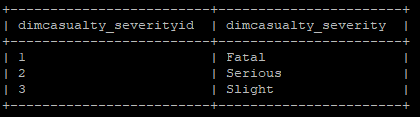
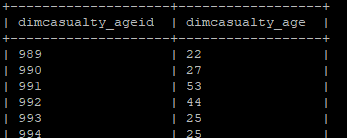
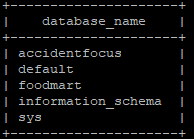
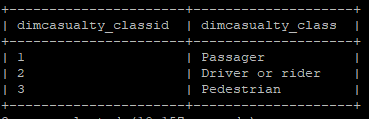
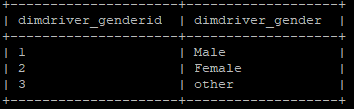
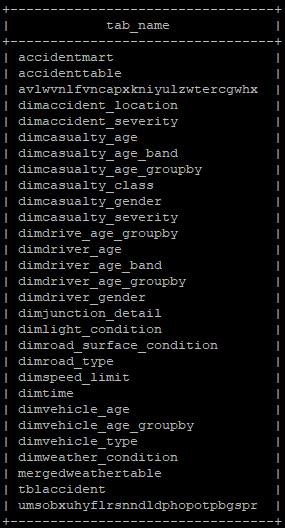
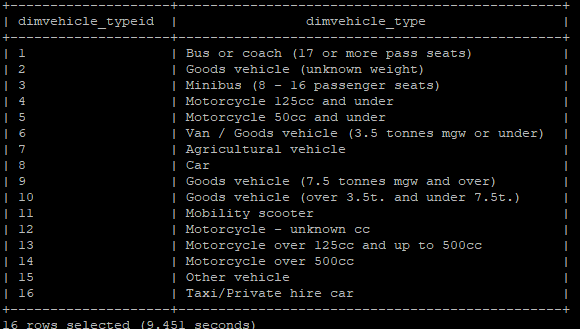
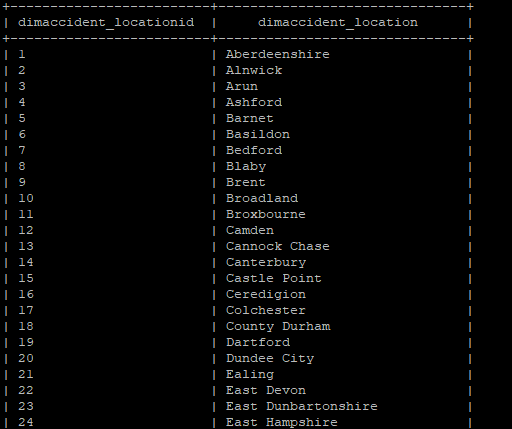
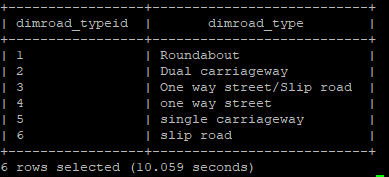
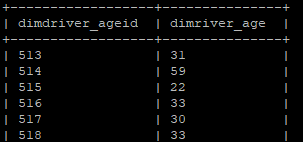
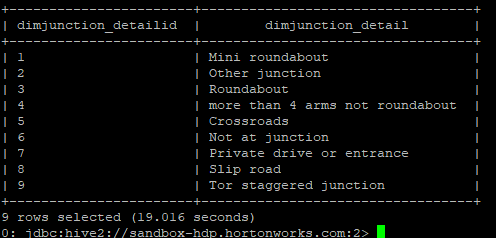
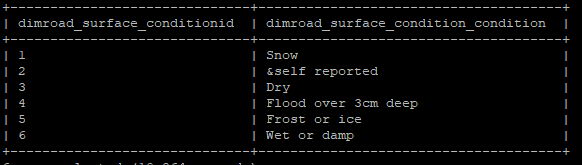
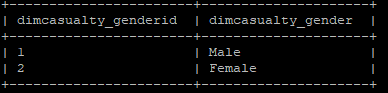
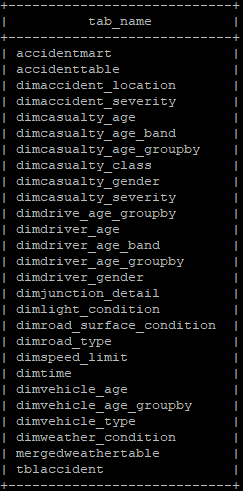
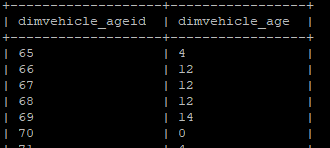


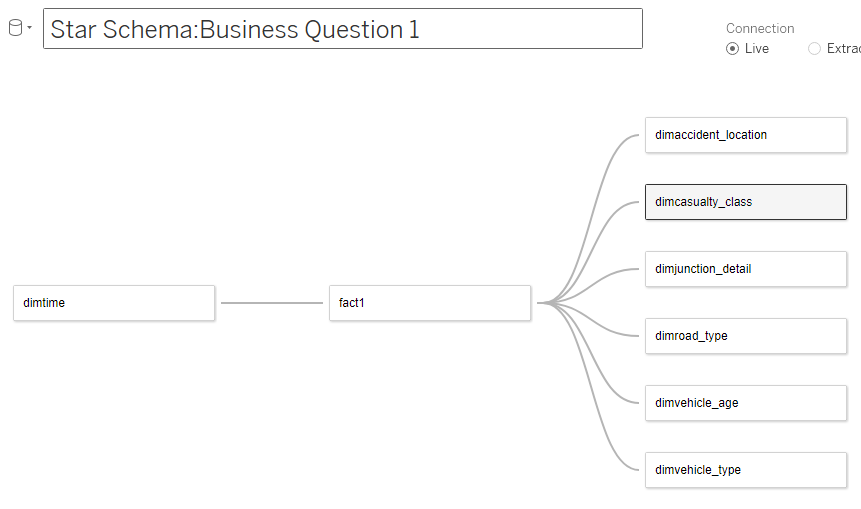








SCREENSHOOT OF ENTITY RELATIONSHIP DIAGRAM



# Appendix

**FIG 1.1: Metadata for our interest variables**

Table

Description automatically generated