Data Exploration

Elliott O'Brien

2021-06-21 15:46:16

Dataset Choices

Primary choice

- The Protest dataset lends itself to many different types of classification and statistical models
- I'm interested to learn from this data, if we can predict if a protest will become violent; to determine what features presented in the dataset are good predictors of violence; comparing multiple models' predictive abilities
- The large number of features (31), would need to be reduced by PCA or a regularization metric such as lasso or ridge for a linear model

```
glimpse(protest)
```

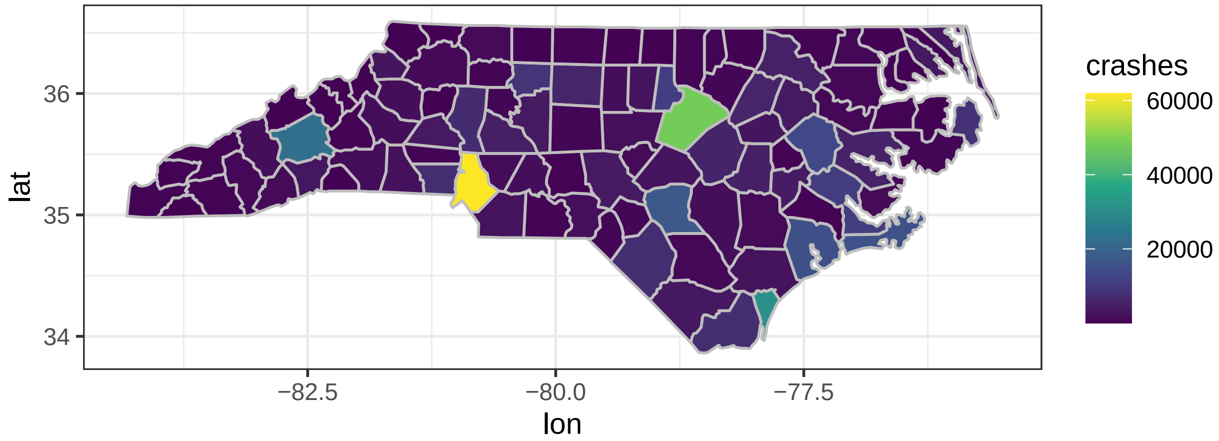
```
## Rows: 17,145
## Columns: 31
## $ id                    <dbl> 201990001, 201990002, 201990003, 201990004, 2019~
## $ country               <chr> "Canada", "Canada", "Canada", "Canada", "Canada"~
## $ ccode                 <dbl> 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, ~
## $ year                  <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1991, 1991, ~
## $ region                <chr> "North America", "North America", "North America~
## $ protest               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ protestnumber         <dbl> 1, 2, 3, 4, 5, 6, 1, 2, 1, 1, 2, 1, 2, 1, 2, 1, ~
## $ startday              <dbl> 15, 25, 1, 12, 14, 19, 10, 28, 4, 16, 1, 1, 18, ~
## $ startmonth            <dbl> 1, 6, 7, 7, 8, 9, 9, 9, 5, 5, 7, 9, 11, 2, 9, 10~
## $ startyear             <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1991, 1991, ~
## $ endday                <dbl> 15, 25, 1, 6, 15, 19, 17, 2, 5, 16, 31, 1, 18, 2~
## $ endmonth              <dbl> 1, 6, 7, 9, 8, 9, 9, 10, 5, 5, 8, 9, 11, 2, 9, 1~
## $ endyear               <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1991, 1991, ~
## $ protesterviolence     <dbl> 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, ~
## $ location              <chr> "national", "Montreal, Quebec", "Montreal, Quebe~
## $ participants_category <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ participants          <chr> "1000s", "1000", "500", "100s", "950", "200", "1~
## $ protesteridentity     <chr> "unspecified", "unspecified", "separatist parti ~
## $ protesterdemand1      <chr> "political behavior, process", "political behavi~
## $ protesterdemand2      <chr> "labor wage dispute", NA, NA, NA, NA, NA, NA, NA~
## $ protesterdemand3      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ protesterdemand4      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ stateresponse1        <chr> "ignore", "ignore", "ignore", "accomodation", "c~
## $ stateresponse2        <chr> NA, NA, NA, NA, "arrests", "shootings", NA, NA, ~
## $ stateresponse3        <chr> NA, NA, NA, NA, "accomodation", NA, NA, NA, NA, ~
## $ stateresponse4        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ stateresponse5        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ stateresponse6        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ stateresponse7        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ sources               <chr> "1. great canadian train journeys into history; ~
## $ notes                 <chr> "canada s railway passenger system was finally c~
```

```
# protest %>%
#   group_by(country) %>%
#   summarize(
#     n_violent_protests = sum(protesterviolence, na.rm = T),
#     n_protests = n(),
#     percent_violent_country = round(n_violent_protests / n() * 100, 2)
#   ) %>%
#   mutate(
#     percent_violent_world = round(n_violent_protests / sum(n_violent_protests) * 100, 2)
#   ) %>%
#   arrange(
#     desc(percent_violent_world)
#   )
```

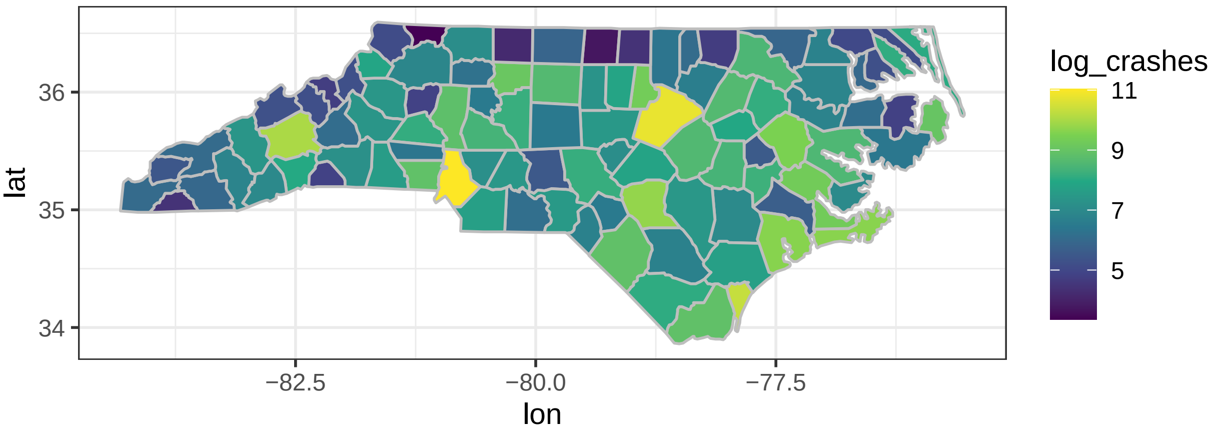
Secondary choice

- The bike_crashes dataset has the most features of the other potential datasets of interest
- data could be visualized on a geographical map
- Models could be developed to classify crash severity, intersection and street with high risk/danger of bike crashes
- Subgroup analyses could be done on types of crashes, severity of injury, etc
- The large number of features (61), would need to be reduced by PCA or a regularization metric such as lasso or ridge for a linear model

```
county_bike_crashes <- bike_crashes %>%
  group_by(County, CrashGrp) %>%
  summarize(
    crashes = n_distinct(CrashID)
  )

NC_map_coords <- map_data("county", "North Carolina") %>%
  transmute(
    lon = long,
    lat,
    group,
    county = str_to_title(subregion)
  )

NC_bike_crash_map_coords <- NC_map_coords %>%
  left_join(
    county_bike_crashes,
    by = c('county' = 'County')
  ) %>%
  group_by(county, ) %>%
  mutate(
    crashes = sum(crashes, na.rm=TRUE),
    log_crashes = log(crashes)
  )

NC_bike_crash_map_coords %>%
  ggplot(aes(lon, lat, group = county, fill = crashes)) +
  geom_polygon(col = 'grey') +
  scale_fill_continuous(type = "viridis") +
  coord_quickmap()
```



```
NC_bike_crash_map_coords %>%
  ggplot(aes(lon, lat, group = county, fill = log_crashes)) +
  geom_polygon(col = 'grey') +
  scale_fill_continuous(type = "viridis") +
  coord_quickmap()
```



```
###########################################################
### Exploring usability of variables present in data ###
###########################################################

## Variables to remove: ##
# * X and Y are longitude and latitude which are already present in the data
# * OBJECTID and OBJECTID_1 seem like a row number id which is not a very useful
#   feature.  Also these are duplicated columns.
# * BikeInjury, DrvrInjury and CrashSevr use similar categories but different
#   values in each record.  These columns will need to be observed closely.
#   since crash severity column has 7 observations where crash_sevr = 'Killed'
#   but neither BikeInjury nor DrvrInjury were considered fatal.
#   Therefore, CrashSevr will be not be counted on.

# # 271 deaths that were biking or driving, mostly bicyclist (269) deaths
# bike_crashes %>%
#   filter(str_detect(BikeInjury, 'K:') | str_detect(DrvrInjury, 'K:'))
# # 2 driver deaths in the whole dataset
# bike_crashes %>%
#   filter(str_detect(DrvrInjury, 'K:'))
# # 278 total deaths, 7 deaths that aren't clear since there is no indication if
# # bicyclist or driver was killed.
# bike_crashes %>%
#   filter(str_detect(CrashSevr, 'K:'))
# # These are the 7 deaths that can't be explained by bicyclist(s) and
# # driver(s). There seems to be a third party involved that is not listed in
# # records, i.e. other pedestrian or other cyclist that didn't cause the crash
# # but was affected by the crash.  It may be good to remove these records since
# # they are misleading without more information on how the deaths occurred.
# bike_crashes %>%
#   filter(
#     str_detect(CrashSevr, 'K:') &
#     !str_detect(BikeInjury, 'K:') &
#     !str_detect(DrvrInjury, 'K:')
#    )

## Duplicates: ##
# Note duplicate crash report for CrashID == 1041566349; note that only
# CrashLoc and CrashType are different. Based on the first record, the driver is
# taking a right turn, but listed as non-intersection; in second record,
# CrashType is "motorist overtaking - Other/Unknown" and Crash Location is
# listed as intersection. Therefore, the CrashLoc was likely an intersection
# where a motorist overtook a bicyclist.  This crash record will need to be
# fixed to reflect this and remove duplication.
bike_crashes[duplicated(bike_crashes$CrashID), 'CrashID'] %>%
  kable(caption = 'Duplicate Record CrashID')
```

Table 1: Duplicate Record CrashID

CrashID
104156349

```
bike_crashes %>%
  filter(CrashID == 104156349) %>%
  glimpse()
```

```
## Rows: 2
## Columns: 62
## $ X          <dbl> -81.20297, -81.20297
## $ Y          <dbl> 35.27395, 35.27395
## $ OBJECTID_1 <dbl> 6901, 6902
## $ AmbulanceR <chr> "No", "No"
## $ BikeAge    <chr> "14", "14"
## $ BikeAgeGrp <chr> "Nov-15", "Nov-15"
## $ BikeAlcDrg <chr> "No", "No"
## $ BikeAlcFlg <chr> "No", "No"
## $ BikeDir    <chr> "With Traffic", "With Traffic"
## $ BikeInjury <chr> "B: Suspected Minor Injury", "B: Suspected Minor Injury"
## $ BikePos    <chr> "Travel Lane", "Travel Lane"
## $ BikeRace   <chr> "Other", "Other"
## $ BikeSex    <chr> "Male", "Male"
## $ City       <chr> "Gastonia", "Gastonia"
## $ County     <chr> "Gaston", "Gaston"
## $ CrashAlcoh <chr> "No", "No"
## $ CrashDay   <chr> "Wednesday", "Wednesday"
## $ CrashGrp   <chr> "Motorist Right Turn / Merge", "Motorist Overtaking Bicycli~
## $ CrashHour  <dbl> 17, 17
## $ CrashID    <dbl> 104156349, 104156349
## $ CrashLoc   <chr> "Non-Intersection", "Intersection-Related"
## $ CrashMonth <chr> "August", "August"
## $ CrashSevr  <chr> "B: Suspected Minor Injury", "B: Suspected Minor Injury"
## $ CrashType  <chr> "Motorist Right Turn - Same Direction", "Motorist Overtakin~
## $ CrashYear  <dbl> 2014, 2014
## $ Developmen <chr> "Residential", "Residential"
## $ DrvrAge    <chr> "999", "999"
## $ DrvrAgeGrp <chr> "Unknown", "Unknown"
## $ DrvrAlcDrg <chr> "Missing", "Missing"
## $ DrvrAlcFlg <chr> "Missing", "Missing"
## $ DrvrInjury <chr> "Unknown Injury", "Unknown Injury"
## $ DrvrRace   <chr> "Unknown/Missing", "Unknown/Missing"
## $ DrvrSex    <chr> "Unknown", "Unknown"
## $ DrvrVehTyp <chr> "Unknown", "Unknown"
## $ HitRun     <chr> "Yes", "Yes"
## $ Latitude   <dbl> 35.27395, 35.27395
## $ LightCond  <chr> "Daylight", "Daylight"
## $ Locality   <chr> "Urban (>70% Developed)", "Urban (>70% Developed)"
## $ Longitude  <dbl> -81.20297, -81.20297
## $ NumBicsAin <chr> "0", "0"
## $ NumBicsBin <chr> "1", "1"
## $ NumBicsCin <chr> "0", "0"
## $ NumBicsKil <chr> "0", "0"
## $ NumBicsNoi <chr> "0", "0"
## $ NumBicsTot <chr> "1", "1"
## $ NumBicsUin <chr> "0", "0"
## $ NumLanes   <chr> "2 lanes", "2 lanes"
## $ NumUnits   <dbl> 2, 2
## $ RdCharacte <chr> "Straight - Grade", "Straight - Grade"
## $ RdClass    <chr> "Local Street", "Local Street"
## $ RdConditio <chr> "Dry", "Dry"
## $ RdConfig   <chr> "Two-Way, Not Divided", "Two-Way, Not Divided"
## $ RdDefects  <chr> "None", "None"
## $ RdFeature  <chr> "Missing", "Missing"
## $ RdSurface  <chr> "Smooth Asphalt", "Smooth Asphalt"
## $ Region     <chr> "Piedmont", "Piedmont"
## $ RuralUrban <chr> "Urban", "Urban"
## $ SpeedLimit <chr> "30 - 35  MPH", "30 - 35  MPH"
## $ TraffCntrl <chr> "No Control Present", "No Control Present"
## $ Weather    <chr> "Clear", "Clear"
## $ Workzone   <chr> "No", "No"
## $ OBJECTID   <dbl> 6901, 6902
```

```
# Fix duplicate record
bike_crashes <-
  bike_crashes %>%
  filter(CrashID != '104156349') %>%
  bind_rows(
    bike_crashes %>%
      filter(CrashID == '104156349') %>%
      slice(1) %>% # grab first record and make changes
      mutate(
        CrashLoc = 'Intersection-Related',
        CrashType = 'Motorist Right Turn - Same Direction',
        CrashGrp = 'Motorist Right Turn / Merge'
      )
  )

glimpse(bike_crashes)
```

```
## Rows: 12,172
## Columns: 62
## $ X          <dbl> -78.88390, -78.78280, -80.69782, -80.47932, -78.90445, -80.~
## $ Y          <dbl> 36.03949, 35.75112, 35.08473, 35.68440, 34.99943, 35.66667,~
## $ OBJECTID_1 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ AmbulanceR <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", "Yes", "Yes~
## $ BikeAge    <chr> "11", "20", "37", "30", "45", "58", "51", "13", "18", "39",~
## $ BikeAgeGrp <chr> "Nov-15", "20-24", "30-39", "30-39", "40-49", "50-59", "50-~
## $ BikeAlcDrg <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ BikeAlcFlg <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No", "No"~
## $ BikeDir    <chr> "With Traffic", "Facing Traffic", "Unknown", "With Traffic"~
## $ BikeInjury <chr> "B: Suspected Minor Injury", "C: Possible Injury", "B: Susp~
## $ BikePos    <chr> "Sidewalk / Crosswalk / Driveway Crossing", "Sidewalk / Cro~
## $ BikeRace   <chr> "Black", "Hispanic", "Black", "White", "Black", "White", "B~
## $ BikeSex    <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Ma~
## $ City       <chr> "Durham", "Cary", "Stallings", "Salisbury", "Fayetteville",~
## $ County     <chr> "Durham", "Wake", "Union", "Rowan", "Cumberland", "Rowan", ~
## $ CrashAlcoh <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No", "No"~
## $ CrashDay   <chr> "Tuesday", "Friday", "Monday", "Friday", "Friday", "Wednesd~
## $ CrashGrp   <chr> "Parallel Paths - Other Circumstances", "Motorist Failed to~
## $ CrashHour  <dbl> 16, 9, 17, 17, 12, 9, 19, 15, 8, 9, 21, 11, 20, 15, 14, 8, ~
## $ CrashID    <dbl> 101878313, 101885911, 101886055, 101890155, 101899756, 1019~
## $ CrashLoc   <chr> "Non-Intersection", "Intersection", "Non-Roadway", "Interse~
## $ CrashMonth <chr> "January", "January", "January", "January", "January", "Jan~
## $ CrashSevr  <chr> "B: Suspected Minor Injury", "C: Possible Injury", "B: Susp~
## $ CrashType  <chr> "Bicyclist Ride Out - Parallel Path", "Motorist Drive Out -~
## $ CrashYear  <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007,~
## $ Developmen <chr> "Residential", "Residential", "Commercial", "Commercial", "~
## $ DrvrAge    <chr> "35", "64", "39", "999", "51", "999", "61", "18", "999", "7~
## $ DrvrAgeGrp <chr> "30-39", "60-69", "30-39", "Unknown", "50-59", "Unknown", "~
## $ DrvrAlcDrg <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ DrvrAlcFlg <chr> "No", "No", "No", "No", "No", "No", "No", "No", "Missing", ~
## $ DrvrInjury <chr> "O: No Injury", "O: No Injury", "O: No Injury", "Unknown In~
## $ DrvrRace   <chr> "White", "White", "White", "Unknown/Missing", "Black", "Unk~
## $ DrvrSex    <chr> "Male", "Male", "Female", "Unknown", "Female", "Unknown", "~
## $ DrvrVehTyp <chr> "Passenger Car", "Passenger Car", "Passenger Car", "Sport U~
## $ HitRun     <chr> "No", "No", "No", "Yes", "No", "Yes", "No", "No", "Yes", "Y~
## $ Latitude   <dbl> 36.03949, 35.75112, 35.08473, 35.68440, 34.99943, 35.66667,~
## $ LightCond  <chr> "Daylight", "Daylight", "Dusk", "Daylight", "Daylight", "Da~
## $ Locality   <chr> "Urban (>70% Developed)", "Urban (>70% Developed)", "Urban ~
## $ Longitude  <dbl> -78.88390, -78.78280, -80.69782, -80.47932, -78.90445, -80.~
## $ NumBicsAin <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ NumBicsBin <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ NumBicsCin <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ NumBicsKil <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ NumBicsNoi <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ NumBicsTot <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ NumBicsUin <chr> ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".", ".",~
## $ NumLanes   <chr> "1 lane", "3 lanes", "2 lanes", "2 lanes", "2 lanes", "2 la~
## $ NumUnits   <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ RdCharacte <chr> "Straight - Level", "Straight - Grade", "Straight - Level",~
## $ RdClass    <chr> "Local Street", "Local Street", "Public Vehicular Area", "L~
## $ RdConditio <chr> "Dry", "Dry", "Dry", "Dry", "Dry", "Dry", "Dry", "Dry", "Dr~
## $ RdConfig   <chr> "Two-Way, Divided, Unprotected Median", "Two-Way, Divided, ~
## $ RdDefects  <chr> "None", "None", "None", "None", "None", "None", "None", "No~
## $ RdFeature  <chr> "No Special Feature", "Four-Way Intersection", "No Special ~
## $ RdSurface  <chr> "Smooth Asphalt", "Smooth Asphalt", "Smooth Asphalt", "Smoo~
## $ Region     <chr> "Piedmont", "Piedmont", "Piedmont", "Piedmont", "Coastal", ~
## $ RuralUrban <chr> "Urban", "Urban", "Urban", "Urban", "Urban", "Urban", "Rura~
## $ SpeedLimit <chr> "30 - 35  MPH", "30 - 35  MPH", "20 - 25  MPH", "30 - 35  M~
## $ TraffCntrl <chr> "No Control Present", "Stop And Go Signal", "No Control Pre~
## $ Weather    <chr> "Clear", "Clear", "Cloudy", "Cloudy", "Clear", "Clear", "Cl~
## $ Workzone   <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No",~
## $ OBJECTID   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
```

Data Preparation

```
bike_crashes_clean <-
  bike_crashes %>%
  select(-X, -Y, -OBJECTID_1, -OBJECTID) %>%
  mutate(
    BikeAge = as.numeric(BikeAge),
    # unlikely anyone over 100 years old
    BikeAge = if_else(BikeAge > 100 | BikeAge == 'Unknown', NA_real_, BikeAge),
    BikeAgeGrp = case_when(
      BikeAgeGrp == '06-Oct' ~ '6-10',
      BikeAgeGrp == 'Nov-15' ~ '11-15',
      BikeAgeGrp == 'Unknown' ~ NA_character_,
      TRUE ~ BikeAgeGrp
    ),
    BikeAlcDrg = case_when(
      BikeAlcDrg == '.' ~ NA_character_,
      BikeAlcDrg == 'Missing' ~ NA_character_,
      BikeAlcDrg == 'Unknown' ~ NA_character_,
      TRUE ~ BikeAlcDrg
    ),
    BikeAlcFlg = case_when(
      BikeAlcFlg == 'Missing' ~ NA_character_,
      BikeAlcFlg == 'Unknown' ~ NA_character_,
      TRUE ~ BikeAlcFlg
    ),
    BikeDir = if_else(BikeDir == 'Unknown', NA_character_, BikeDir),
    BikePos = if_else(BikePos == 'Unknown', NA_character_, BikePos),
    BikeRace = if_else(BikeRace == 'Unknown/Missing', NA_character_, BikeRace),
    BikeSex = if_else(BikeSex == 'Unknown', NA_character_, BikeSex),
    CrashGrp = if_else(CrashGrp == 'Other / Unknown - Insufficient Details', NA_character_, CrashGrp),
    CrashLoc = if_else(CrashLoc == 'Unknown Location', NA_character_, CrashLoc),
    CrashDay = factor(CrashDay,
      levels = c('Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday')
    ),
    CrashMonth = factor(CrashMonth,
      levels = c('January', 'February', 'March' ,'April', 'May', 'June', 'July', 'August', 'September',
    ),
    CrashType = if_else(CrashType %in% c('Unknown Approach Paths', 'Unknown Location'), NA_character_, C
    DrvrAge = as.numeric(DrvrAge),
    # unlikely anyone over 100 years old
    DrvrAge = if_else(DrvrAge > 100 | DrvrAge == '70+' | DrvrAge == 'Unknown', NA_real_, DrvrAge),
    DrvrAgeGrp = if_else(DrvrAgeGrp == 'Unknown', NA_character_, DrvrAgeGrp),
    DrvrAlcDrg = case_when(
      DrvrAlcDrg == '.' ~ NA_character_,
      DrvrAlcDrg == 'Missing' ~ NA_character_,
      DrvrAlcDrg == 'Unknown' ~ NA_character_,
      TRUE ~ DrvrAlcDrg
    ),
    DrvrAlcFlg = case_when(
      DrvrAlcFlg == 'Missing' ~ NA_character_,
      DrvrAlcFlg == 'Unknown' ~ NA_character_,
      TRUE ~ DrvrAlcFlg
    ),
    DrvrRace = if_else(DrvrRace == 'Unknown/Missing', NA_character_, DrvrRace),
    DrvrSex = if_else(DrvrSex == 'Unknown', NA_character_, DrvrSex),
    LightCond = if_else(LightCond == 'Unknown', NA_character_, LightCond),
    NumBicsAin = if_else(NumBicsAin == '.', NA_character_, NumBicsAin),
    NumBicsBin = if_else(NumBicsBin == '.', NA_character_, NumBicsBin),
    NumBicsCin = if_else(NumBicsCin == '.', NA_character_, NumBicsCin),
    NumBicsKil = if_else(NumBicsKil == '.', NA_character_, NumBicsKil),
    NumBicsNoi = if_else(NumBicsNoi == '.', NA_character_, NumBicsNoi),
    NumBicsUin = if_else(NumBicsUin == '.', NA_character_, NumBicsUin),
    NumBicsTot = if_else(NumBicsTot == '.', NA_character_, NumBicsTot),
    NumLanes = if_else(NumLanes == 'Unknown', NA_character_, NumLanes),
    NumLanes = if_else(NumLanes == '9 or more lanes', '9+ lanes', NumLanes),
    NumLanes = factor(NumLanes,
      levels = c('1 lane', '2 lanes', '3 lanes', '4 lanes', '5 lanes', '6 lanes', '7 lanes', '8 lanes',
    ),
    RdCharacte = if_else(RdCharacte == 'Unknown', NA_character_, RdCharacte),
    RdClass = if_else(RdClass %in% c('.', 'missing', 'Unknown'), NA_character_, RdClass),
    RdConditio = if_else(RdConditio == 'Unknown', NA_character_, RdConditio),
    RdConfig = if_else(RdConfig == 'Unknown', NA_character_, RdConfig),
    RdDefects = if_else(RdDefects %in% c('Unknown', 'Missing'), NA_character_, RdDefects),
    RdFeature = if_else(RdFeature == 'Missing', NA_character_, RdFeature),
    RdSurface = if_else(RdSurface %in% c('Unknown', 'Missing'), NA_character_, RdSurface),
    RuralUrban = if_else(RuralUrban == '.', NA_character_, RuralUrban),
    SpeedLimit = if_else(SpeedLimit == 'Unknown', NA_character_, SpeedLimit),
    SpeedLimit = str_replace(SpeedLimit, '\\s{2}', ' '), # extra spaces between meaure and units have b
    SpeedLimit = factor(SpeedLimit,
      levels = c('5 - 15 MPH', '20 - 25 MPH', '30 - 35 MPH', '40 - 45 MPH', '50 - 55 MPH', '60 - 75 MPH
    ),
    TraffCntrl = if_else(TraffCntrl == 'Missing', NA_character_, TraffCntrl)
  ) %>%
  # transform char columns that need to be split
  separate(
    col = BikeInjury,
    into = c('BikeInjuryCat', 'BikeInjuryDisc'),
    sep = ': '
  ) %>%
  separate(
    col = CrashSevr,
    into = c('CrashSevrCat', 'CrashSevrDisc'),
    sep = ': '
  ) %>%
  separate(
    col = DrvrInjury,
    into = c('DrvrInjuryCat', 'DrvrInjuryDisc'),
    sep = ': '
  ) %>%
  mutate(
    BikeInjuryCat = if_else(BikeInjuryCat == 'Unknown Injury', 'U', BikeInjuryCat),
    BikeInjuryDisc = if_else(is.na(BikeInjuryDisc), 'Unknown Injury', BikeInjuryDisc),
    CrashSevrCat = if_else(CrashSevrCat == 'Unknown Injury', 'U', CrashSevrCat),
    CrashSevrDisc = if_else(is.na(CrashSevrDisc), 'Unknown Injury', CrashSevrDisc),
    DrvrInjuryCat = if_else(DrvrInjuryCat == 'Unknown Injury', 'U', DrvrInjuryCat),
    DrvrInjuryDisc = if_else(is.na(DrvrInjuryDisc), 'Unknown Injury', DrvrInjuryDisc)
  ) %>%
  # group specific calculations and imputations
  group_by(BikeAgeGrp) %>%
  mutate(
    # median imputation for age group 70+
    BikeAge = if_else(is.na(BikeAge), median(BikeAge, na.rm = T), BikeAge)
  ) %>%
  ungroup() %>%
  group_by(DrvrAgeGrp) %>%
  mutate(
    # median imputation for age (years) recorded as 70+
    DrvrAge = if_else(is.na(DrvrAge), median(DrvrAge, na.rm = T), DrvrAge)
  ) %>%
  ungroup() %>%
  mutate(
    across(is.character, ~ as.factor(.x))
  )

glimpse(bike_crashes_clean)
```

```
## Rows: 12,172
## Columns: 61
## $ AmbulanceR     <fct> Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, No, Yes, Ye~
## $ BikeAge        <dbl> 11, 20, 37, 30, 45, 58, 51, 13, 18, 39, 19, 35, 40, 31,~
## $ BikeAgeGrp     <fct> 11-15, 20-24, 30-39, 30-39, 40-49, 50-59, 50-59, 11-15,~
## $ BikeAlcDrg     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ BikeAlcFlg     <fct> No, No, No, No, No, No, Yes, No, No, No, Yes, No, No, N~
## $ BikeDir        <fct> With Traffic, Facing Traffic, NA, With Traffic, With Tr~
## $ BikeInjuryCat  <fct> B, C, B, C, B, B, A, C, C, B, A, B, C, B, C, C, A, C, C~
## $ BikeInjuryDisc <fct> Suspected Minor Injury, Possible Injury, Suspected Mino~
## $ BikePos        <fct> Sidewalk / Crosswalk / Driveway Crossing, Sidewalk / Cr~
## $ BikeRace       <fct> Black, Hispanic, Black, White, Black, White, Black, Whi~
## $ BikeSex        <fct> Male, Male, Male, Male, Male, Male, Male, Male, Male, M~
## $ City           <fct> Durham, Cary, Stallings, Salisbury, Fayetteville, Salis~
## $ County         <fct> Durham, Wake, Union, Rowan, Cumberland, Rowan, Randolph~
## $ CrashAlcoh     <fct> No, No, No, No, No, No, Yes, No, No, No, Yes, No, No, N~
## $ CrashDay       <fct> Tuesday, Friday, Monday, Friday, Friday, Wednesday, Sat~
## $ CrashGrp       <fct> Parallel Paths - Other Circumstances, Motorist Failed t~
## $ CrashHour      <dbl> 16, 9, 17, 17, 12, 9, 19, 15, 8, 9, 21, 11, 20, 15, 14,~
## $ CrashID        <dbl> 101878313, 101885911, 101886055, 101890155, 101899756, ~
## $ CrashLoc       <fct> Non-Intersection, Intersection, Non-Roadway, Intersecti~
## $ CrashMonth     <fct> January, January, January, January, January, January, J~
## $ CrashSevrCat   <fct> B, C, B, C, B, B, A, C, C, B, A, B, C, B, C, C, A, C, C~
## $ CrashSevrDisc  <fct> Suspected Minor Injury, Possible Injury, Suspected Mino~
## $ CrashType      <fct> "Bicyclist Ride Out - Parallel Path", "Motorist Drive O~
## $ CrashYear      <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2~
## $ Developmen     <fct> "Residential", "Residential", "Commercial", "Commercial~
## $ DrvrAge        <dbl> 35, 64, 39, NA, 51, NA, 61, 18, NA, 76, 24, 27, 21, 17,~
## $ DrvrAgeGrp     <fct> 30-39, 60-69, 30-39, NA, 50-59, NA, 60-69, 0-19, NA, 70~
## $ DrvrAlcDrg     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ DrvrAlcFlg     <fct> No, No, No, No, No, No, No, No, NA, No, Yes, No, No, No~
## $ DrvrInjuryCat  <fct> O, O, O, U, O, U, O, O, U, U, O, O, O, O, O, O, O, O, O~
## $ DrvrInjuryDisc <fct> No Injury, No Injury, No Injury, Unknown Injury, No Inj~
## $ DrvrRace       <fct> White, White, White, NA, Black, NA, White, Black, NA, N~
## $ DrvrSex        <fct> Male, Male, Female, NA, Female, NA, Male, Female, NA, M~
## $ DrvrVehTyp     <fct> "Passenger Car", "Passenger Car", "Passenger Car", "Spo~
## $ HitRun         <fct> No, No, No, Yes, No, Yes, No, No, Yes, Yes, No, No, No,~
## $ Latitude       <dbl> 36.03949, 35.75112, 35.08473, 35.68440, 34.99943, 35.66~
## $ LightCond      <fct> Daylight, Daylight, Dusk, Daylight, Daylight, Daylight,~
## $ Locality       <fct> Urban (>70% Developed), Urban (>70% Developed), Urban (~
## $ Longitude      <dbl> -78.88390, -78.78280, -80.69782, -80.47932, -78.90445, ~
## $ NumBicsAin     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ NumBicsBin     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ NumBicsCin     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ NumBicsKil     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ NumBicsNoi     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ NumBicsTot     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ NumBicsUin     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ NumLanes       <fct> 1 lane, 3 lanes, 2 lanes, 2 lanes, 2 lanes, 2 lanes, 2 ~
## $ NumUnits       <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ RdCharacte     <fct> Straight - Level, Straight - Grade, Straight - Level, S~
## $ RdClass        <fct> "Local Street", "Local Street", "Public Vehicular Area"~
## $ RdConditio     <fct> "Dry", "Dry", "Dry", "Dry", "Dry", "Dry", "Dry", "Dry",~
## $ RdConfig       <fct> "Two-Way, Divided, Unprotected Median", "Two-Way, Divid~
## $ RdDefects      <fct> "None", "None", "None", "None", "None", "None", "None",~
## $ RdFeature      <fct> "No Special Feature", "Four-Way Intersection", "No Spec~
## $ RdSurface      <fct> Smooth Asphalt, Smooth Asphalt, Smooth Asphalt, Smooth ~
## $ Region         <fct> Piedmont, Piedmont, Piedmont, Piedmont, Coastal, Piedmo~
## $ RuralUrban     <fct> Urban, Urban, Urban, Urban, Urban, Urban, Rural, Rural,~
## $ SpeedLimit     <fct> 30 - 35 MPH, 30 - 35 MPH, 20 - 25 MPH, 30 - 35 MPH, 30 ~
## $ TraffCntrl     <fct> "No Control Present", "Stop And Go Signal", "No Control~
## $ Weather        <fct> "Clear", "Clear", "Cloudy", "Cloudy", "Clear", "Clear",~
## $ Workzone       <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No,~
```

```r
saveRDS(bike_crashes_clean, './derived_data/bike_crashes_clean.rds')
```