Data Pre-processing

Zhuocheng Lin

3/18/2020

Packages

```
library(tidyverse)
library(modelr)
library(lubridate)
library(caret)
```

Read data

```
df <- read_csv('./US_Accidents_Dec19.csv', col_types = cols(.default = col_character())) %>%
  type_convert()
```

Drop variables with high NA proportion (over 50%)

```
df %>% summarise_all(~ mean(is.na(.))) %>%
  pivot_longer(1:49, names_to = "variable", values_to = "NA_prop") %>%
  filter(NA_prop >= 0.5)
```

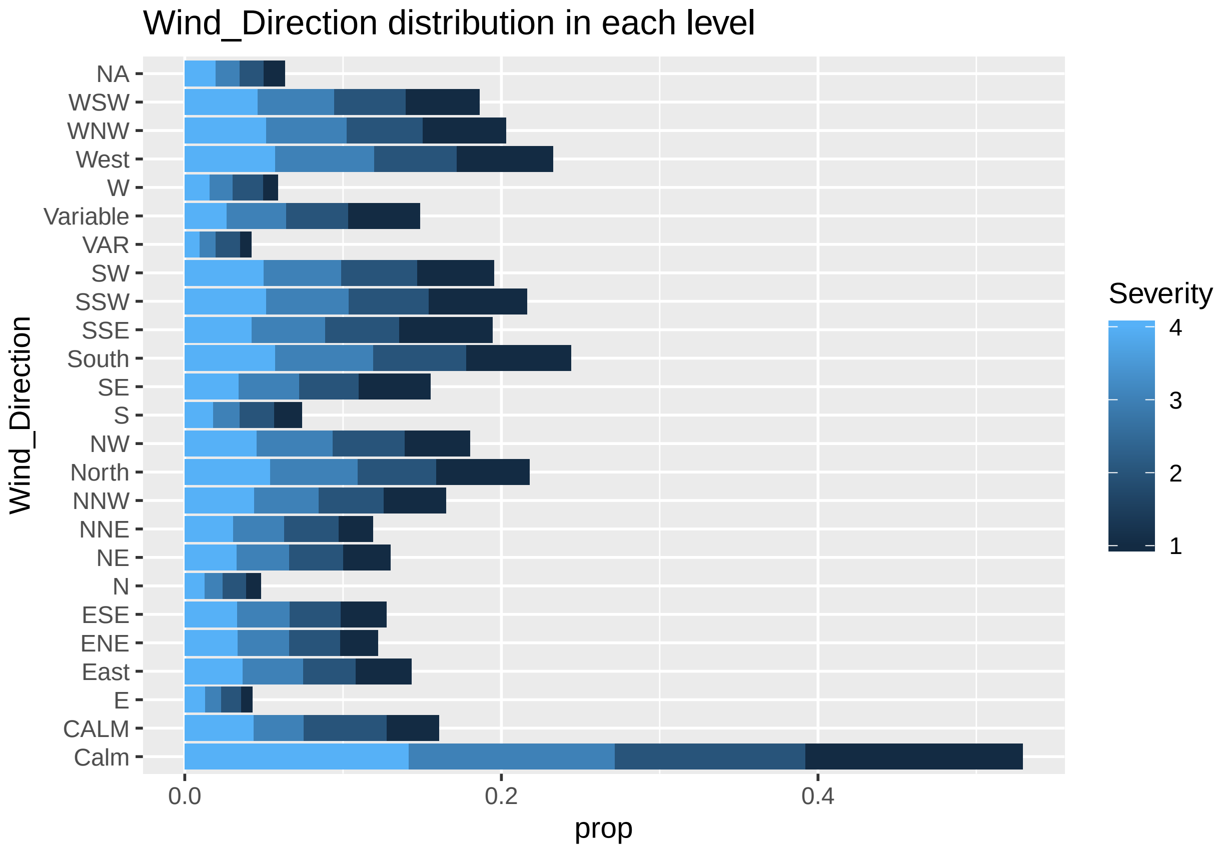
```
## # A tibble: 5 x 2
##   variable          NA_prop
##   <chr>               <dbl>
## 1 End_Lat             0.755
## 2 End_Lng             0.755
## 3 Number              0.645
## 4 Wind_Chill(F)       0.623
## 5 Precipitation(in)   0.672
```

```
drop_na_cols <- c("End_Lat", "End_Lng", "Number", "Wind_Chill(F)", "Precipitation(in)")
```

Drop unuseful variable

```
# these variables are not useful in predicting Severity
not_useful <- c("ID", "Source", "Timezone", "Airport_Code", "Weather_Timestamp",
                "Wind_Direction", "Country", "Description")

# Not so sure about whether Wind_Direction is useful
# to me, the relation seems weak
df %>% ggplot(aes(Wind_Direction, ..prop..)) +
    geom_bar(aes(group = Severity, fill = Severity)) +
    coord_flip() +
    labs(title = "Wind_Direction distribution in each level")
```



```
df_drop <- df %>% select(-drop_na_cols, -not_useful)
```

Rename variables to avoid potential error

```
df_drop <-  df_drop %>%
  rename("Distance" = `Distance(mi)`, "Temperature" = `Temperature(F)`, "Humidity" = `Humidity(%)`,
         "Pressure" = `Pressure(in)`, "Visibility" = `Visibility(mi)`, "Wind_Speed" = `Wind_Speed(mph)`
```

Pre-processing time related variables

```
df_time <- df_drop %>%
  mutate(Duration = as.numeric(End_Time - Start_Time)) %>%
  # accident duration should be positive
  filter(!(Duration < 0)) %>%
  separate(Start_Time, into = c("Date", "Time"), sep = " ") %>%
  mutate("Year" = str_sub(Date, 1, 4), "Month" = str_sub(Date, 6, 7), "Day" = str_sub(Date, 9, 10),
         "Wday" = as.character(wday(Date)), "Hour" = str_sub(Time, 1, 2)) %>%
  select(-c("Date", "Time", "End_Time")) %>%
  select(TMC, Severity, Year, Month, Day, Hour, Wday, Duration, everything())
```

```r
head(df_time)
```

```
## # A tibble: 6 x 40
##     TMC Severity Year  Month Day   Hour  Wday  Duration Start_Lat Start_Lng
##   <dbl>    <dbl> <chr> <chr> <chr> <chr> <chr>    <dbl>     <dbl>     <dbl>
## 1   201        3 2016  02    08    05    2        18840      39.9     -84.1
## 2   201        2 2016  02    08    06    2         1800      39.9     -82.8
## 3   201        2 2016  02    08    06    2         1800      39.1     -84.0
## 4   201        3 2016  02    08    07    2         1800      39.7     -84.2
## 5   201        2 2016  02    08    07    2         1800      39.6     -84.2
## 6   201        3 2016  02    08    07    2         1800      40.1     -82.9
## # ... with 30 more variables: Distance <dbl>, Street <chr>, Side <chr>,
## #   City <chr>, County <chr>, State <chr>, Zipcode <chr>, Temperature <dbl>,
## #   Humidity <dbl>, Pressure <dbl>, Visibility <dbl>, Wind_Speed <dbl>,
## #   Weather_Condition <chr>, Amenity <lgl>, Bump <lgl>, Crossing <lgl>,
## #   Give_Way <lgl>, Junction <lgl>, No_Exit <lgl>, Railway <lgl>,
## #   Roundabout <lgl>, Station <lgl>, Stop <lgl>, Traffic_Calming <lgl>,
## #   Traffic_Signal <lgl>, Turning_Loop <lgl>, Sunrise_Sunset <chr>,
## #   Civil_Twilight <chr>, Nautical_Twilight <chr>, Astronomical_Twilight <chr>
```

Address

```r
# not sure the best way to deal with address
# my opinion is we can choose one state data, and build the model
# and ignore Street, County and City
address <- c("Street", "County", "City", "Zipcode")
df_add <- df_time %>% select(-address)
```

Drop missing Weather_Condition

```r
# when Weather_Condition is missing,
# other variables related to weather will be missing too (most cases)
df_add %>% filter(is.na(Weather_Condition)) %>% select(Temperature:Weather_Condition)
```

```
## # A tibble: 65,932 x 6
##    Temperature Humidity Pressure Visibility Wind_Speed Weather_Condition
##          <dbl>    <dbl>    <dbl>      <dbl>      <dbl> <chr>
##  1        48.2       93     29.5         10        9.2 <NA>
##  2        NA         NA     NA           NA       NA   <NA>
##  3        95         20     29.9         10        6.9 <NA>
##  4        91.4       28     29.9         10       15   <NA>
##  5        NA         NA     NA           NA       NA   <NA>
##  6        NA         NA     NA           NA       NA   <NA>
##  7        NA         NA     NA           NA       NA   <NA>
##  8        NA         NA     NA           NA       NA   <NA>
##  9        NA         NA     NA           NA       NA   <NA>
## 10        NA         NA     NA           NA       NA   <NA>
## # ... with 65,922 more rows
```

```r
df_add %>% filter(is.na(Weather_Condition)) %>% select(Temperature:Weather_Condition) %>%
  summarise_all(~sum(is.na(.)))
```

```
## # A tibble: 1 x 6
##   Temperature Humidity Pressure Visibility Wind_Speed Weather_Condition
##         <int>    <int>    <int>      <int>      <int>             <int>
## 1       46246    46309    44532      58500      56084             65932
```

```
# we can drop observations whose Weather_Condition is missing
df_weather <- df_add %>% filter(!is.na(Weather_Condition))
```

Format

```
df_weather <- df_weather %>%
  mutate(TMC = as.character(TMC)) %>%
  mutate_if(is.logical, as.character)
```

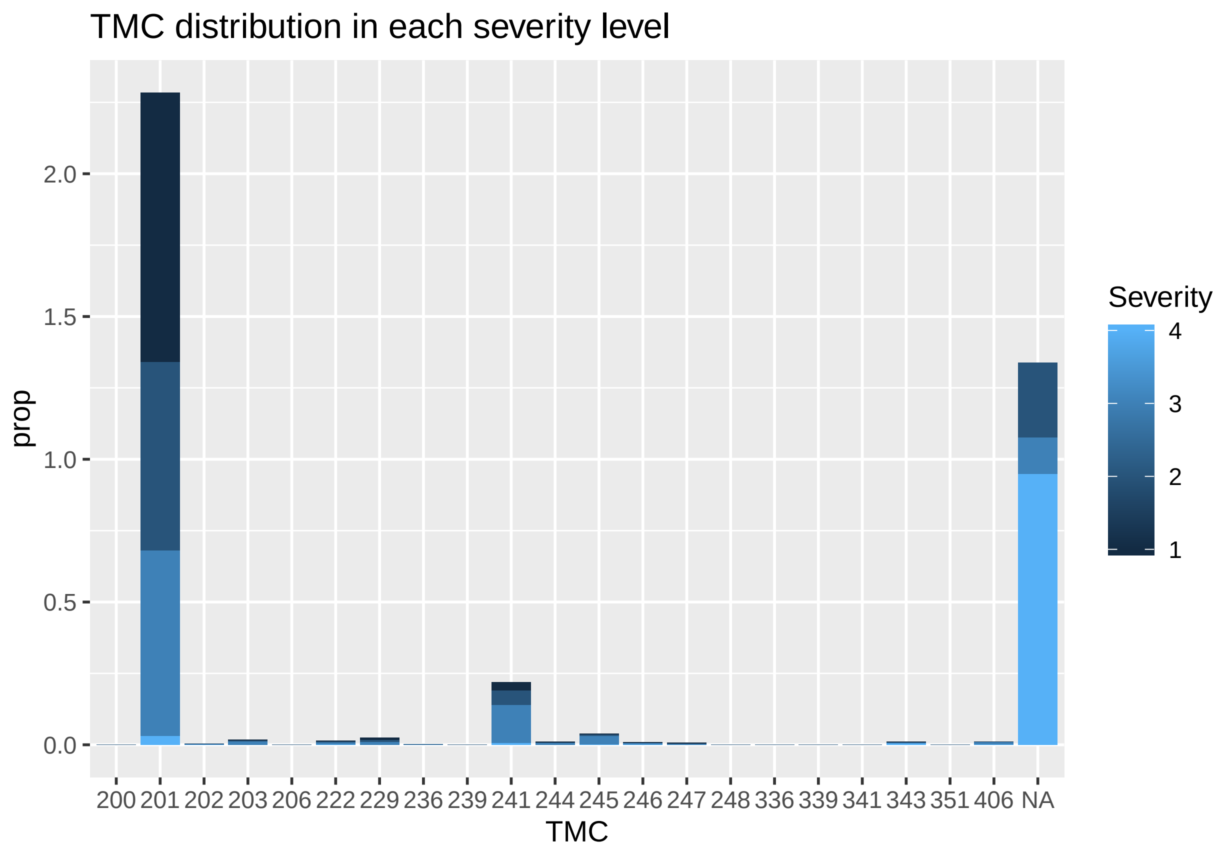
Replace NA with mean

```
df_mean <- df_weather %>%
  mutate_if(is.numeric, ~ replace_na(., mean(., na.rm = T)))
summary(df_mean %>% select_if(is.numeric))
```

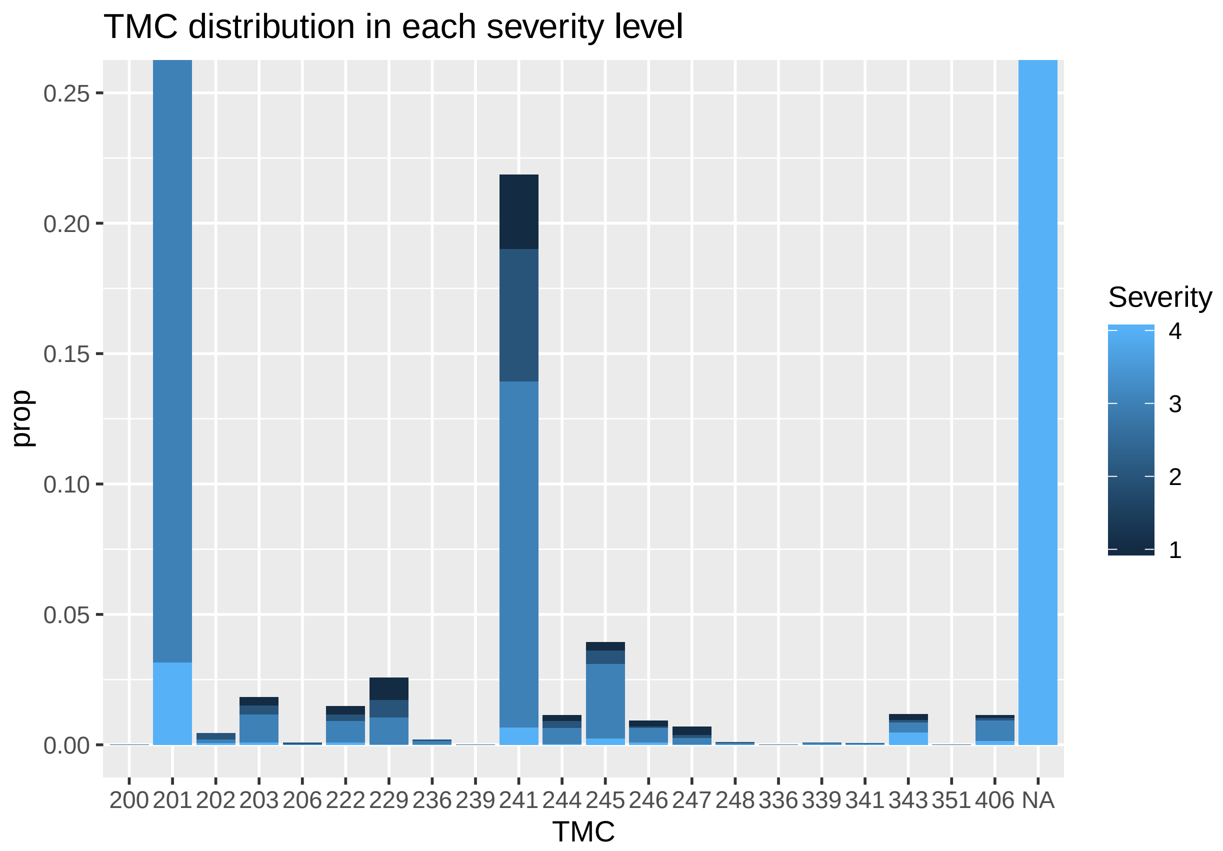
```
##     Severity        Duration          Start_Lat       Start_Lng
##  Min.   :1.000   Min.   :      73   Min.   :24.56   Min.   :-124.62
##  1st Qu.:2.000   1st Qu.:    1783   1st Qu.:33.54   1st Qu.:-117.30
##  Median :2.000   Median :    2675   Median :35.82   Median : -90.25
##  Mean   :2.359   Mean   :    7063   Mean   :36.48   Mean   : -95.47
##  3rd Qu.:3.000   3rd Qu.:    4481   3rd Qu.:40.41   3rd Qu.: -80.95
##  Max.   :4.000   Max.   :91680802   Max.   :49.00   Max.   : -67.11
##     Distance         Temperature        Humidity         Pressure
##  Min.   :  0.0000   Min.   :-40.00   Min.   :  1.00   Min.   : 0.00
##  1st Qu.:  0.0000   1st Qu.: 50.00   1st Qu.: 49.00   1st Qu.:29.82
##  Median :  0.0000   Median : 64.40   Median : 67.00   Median :29.98
##  Mean   :  0.2831   Mean   : 62.38   Mean   : 65.41   Mean   :29.83
##  3rd Qu.:  0.0100   3rd Qu.: 76.00   3rd Qu.: 84.00   3rd Qu.:30.11
##  Max.   :333.6300   Max.   :170.60   Max.   :100.00   Max.   :33.04
##    Visibility        Wind_Speed
##  Min.   :  0.000   Min.   :  0.000
##  1st Qu.: 10.000   1st Qu.:  5.800
##  Median : 10.000   Median :  8.100
##  Mean   :  9.151   Mean   :  8.296
##  3rd Qu.: 10.000   3rd Qu.: 10.400
##  Max.   :140.000   Max.   :822.800
```

TMC

```
# most TMC NAs are in level 4
df_mean %>%
  ggplot(aes(TMC, ..prop..)) +
    geom_bar(aes(group = Severity, fill = Severity)) +
    labs(title = "TMC distribution in each severity level")
```



```
df_mean %>%
  ggplot(aes(TMC, ..prop..)) +
    geom_bar(aes(group = Severity, fill = Severity)) +
    labs(title = "TMC distribution in each severity level") +
    coord_cartesian(ylim = c(0, 0.25))
```



```
# my opinion is TMC NA can be considered as an important feature of Severity
# we can treate NA as a new TMC code
df_TMC <- df_mean %>%
  mutate(TMC = replace_na(TMC, "NA"))
```

Final check if there is unusual observation

```
df_TMC %>% summarise_all(~sum(is.na(.))) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "NAs") %>% filter(NAs > 0)
```

```
## # A tibble: 5 x 2
##   variable                NAs
##   <chr>                 <int>
## 1 Side                      1
## 2 Sunrise_Sunset           80
## 3 Civil_Twilight           80
## 4 Nautical_Twilight        80
## 5 Astronomical_Twilight    80
```

```
# Side has 1 NA, remove it
# variables related to daylight all have 80 NAs

df_TMC %>% filter(is.na(Sunrise_Sunset)) %>% count(TMC)
```

```
## # A tibble: 6 x 2
##   TMC       n
##   <chr> <int>
## 1 201      39
## 2 222       1
## 3 229       2
## 4 241       2
## 5 343       1
## 6 NA       35
```

```
# the missing daylight data may be related to missing TMC
# replace them with a new levle "NAs"
df_final <- df_TMC %>%
  filter(!is.na(Side)) %>%
  filter(!is.na(Sunrise_Sunset))
```

Write csv file

```
# write_csv(df_final, "./tidy.csv")
```