Assignment 2

Jaxon Abercrombie

10/6/2021

Data Wrangling

```
chsInd <- read.csv("chs_individual.csv")
chsReg <- read.csv("chs_regional.csv")
chsCombo <- smartbind(chsInd, chsReg)

# Check location variable to ensure no N/As
sum(is.na(chsCombo$townname))
```

Read in the Data

```
## [1] 0
```

```
# Check dimensions to ensure no duplicates
dim(chsInd)
```

Q1

```
## [1] 1200   23
```

```
dim(chsReg)
```

```
## [1] 12 27
```

```
dim(chsCombo)
```

```
## [1] 1212   49
```

```
# The combo of the two data sets has the sum of the two row values from the dim() function
# To ensure there are not duplicates...
chsCombo <- unique(chsCombo)
chsCombo <- data.table(chsCombo)

# Impute values for those that are missing; to continuous variables only
chsMH <-
  chsCombo %>%
  filter (male == 1 & hispanic == 1)

chsCombo[is.na(agepft), agepft := mean(chsMH$agepft, na.rm = TRUE)]
chsCombo[is.na(height), height := mean(chsMH$height, na.rm = TRUE)]
```

```
## Warning in `[.data.table`(chsCombo, is.na(height), `:=`(height,
## mean(chsMH$height, : 138.598394 (type 'double') at RHS position 1 truncated
## (precision lost) when assigning to type 'integer' (column 7 named 'height')
```

```
chsCombo[is.na(weight), weight := mean(chsMH$weight, na.rm = TRUE)]
```

```
## Warning in `[.data.table`(chsCombo, is.na(weight), `:=`(weight,
## mean(chsMH$weight, : 82.767068 (type 'double') at RHS position 1 truncated
## (precision lost) when assigning to type 'integer' (column 8 named 'weight')
```

```
chsCombo[is.na(bmi), bmi := mean(chsMH$bmi, na.rm = TRUE)]
chsCombo[is.na(fev), fev := mean(chsMH$fev, na.rm = TRUE)]
chsCombo[is.na(fvc), fvc := mean(chsMH$fvc, na.rm = TRUE)]
chsCombo[is.na(mmef), mmef := mean(chsMH$mmef, na.rm = TRUE)]
chsCombo[is.na(educ_parent), educ_parent := mean(chsMH$educ_parent, na.rm = TRUE)]
```

```
## Warning in `[.data.table`(chsCombo, is.na(educ_parent), `:=`(educ_parent, :
## 2.423868 (type 'double') at RHS position 1 truncated (precision lost) when
## assigning to type 'integer' (column 17 named 'educ_parent')
```

```
chsCombo[is.na(male), male := round(mean(chsMH$male, na.rm = TRUE), 1)]
chsCombo[is.na(asthma), asthma := round(mean(chsMH$asthma, na.rm = TRUE), 1)]
```

```
## Warning in `[.data.table`(chsCombo, is.na(asthma), `:=`(asthma,
## round(mean(chsMH$asthma, : 0.200000 (type 'double') at RHS position 1 truncated
## (precision lost) when assigning to type 'integer' (column 10 named 'asthma')
```

```
chsCombo[is.na(smoke), smoke := round(mean(chsMH$smoke, na.rm = TRUE), 1)]
```

```
## Warning in `[.data.table`(chsCombo, is.na(smoke), `:=`(smoke,
## round(mean(chsMH$smoke, : 0.200000 (type 'double') at RHS position 1 truncated
## (precision lost) when assigning to type 'integer' (column 18 named 'smoke')
```

```
chsCombo[is.na(gasstove), gasstove := round(mean(chsMH$educ_parent, na.rm = TRUE), 1)]
```

```
## Warning in `[.data.table`(chsCombo, is.na(gasstove), `:=`(gasstove,
## round(mean(chsMH$educ_parent, : 2.400000 (type 'double') at RHS position 1
## truncated (precision lost) when assigning to type 'integer' (column 20 named
## 'gasstove')
```

```
# BMI; Numerical to categorical
chsCombo$bmiCat <- cut(chsCombo$bmi,
                       breaks = c(0,14,22,24,Inf),
                       labels = c("underweight", "normal", "overweight", "obese"))

# Ensure coding is correct by viewing range of values within each BMI category
chsCombo %>%
  group_by(bmiCat) %>%
  summarize(
    min = min(bmi),
    max = max(bmi),
    count = n()
  )
```

Q2

```
## # A tibble: 4 x 4
##   bmiCat        min   max count
##   <fct>       <dbl> <dbl> <int>
## 1 underweight  11.3  14.0    35
## 2 normal       14.0  22.0   987
```

```
## 3 overweight   22.0  24.0    87
## 4 obese        24.0  41.3   103
```

```
# Some values of gasstove are 2, which would create NAs in our smoke_gas_exposure variable
# This is corrected by adding the "| 2" in our code to allow 1 or 2 to be for gasstove
# Create smoke and gas exposure variable, ensuring the 4 different combinations are taken care of
chsCombo <-
  chsCombo %>%
  mutate(smoke_gas_exposure = case_when(smoke == 0 & gasstove == 0 ~ "neither",
                                        smoke == 1 & gasstove == 0 ~ "smoke only",
                                        smoke == 0 & gasstove == 1 ~ "gas stove only",
                                        smoke == 0 & gasstove == 2 ~ "gas stove only",
                                        smoke == 1 & gasstove == 1 ~ "both",
                                        smoke == 1 & gasstove == 2 ~ "both"))

# Check to see that there are 4 distinct categories and that they match the above code
table(chsCombo$smoke_gas_exposure)
```

Q3

```
##
##           both gas stove only        neither     smoke only
##            154            803            219             36
```

```
sum(is.na(chsCombo$smoke_gas_exposure))
```

```
## [1] 0
```

```
# By town
chsCombo %>%
  group_by(townname) %>%
  summarize(count = n(),
            meanFEV = mean(fev),
            sdFEV = sd(fev),
            percAsthma = 100*mean(asthma, na.rm = TRUE),
            sdAsthma = sd(asthma, na.rm = TRUE))
```

Q4

```
## # A tibble: 12 x 6
##    townname      count meanFEV sdFEV percAsthma sdAsthma
##    <chr>         <int>   <dbl> <dbl>      <dbl>    <dbl>
##  1 Alpine          101   2091.  289.       10.9    0.313
##  2 Atascadero      101   2082.  322.       24.8    0.434
##  3 Lake Elsinore   101   2048.  302.       11.9    0.325
##  4 Lake Gregory    101   2095.  317.       14.9    0.357
##  5 Lancaster       101   2018.  317.       15.8    0.367
##  6 Lompoc          101   2046.  349.       10.9    0.313
##  7 Long Beach      101   1995.  319.       12.9    0.337
##  8 Mira Loma       101   1995.  325.       14.9    0.357
##  9 Riverside       101   1999.  278.       10.9    0.313
## 10 San Dimas       101   2031.  317.       16.8    0.376
## 11 Santa Maria     101   2034.  311.       12.9    0.337
## 12 Upland          101   2036.  342.       11.9    0.325
```

```
# By sex
chsCombo <- chsCombo %>%
  mutate(sex = factor(male,
                      levels = c(0,1),
                      labels = c("male","female")))

chsCombo %>%
  group_by(sex) %>%
  summarize(count = n(),
          meanFEV = mean(fev),
          sdFEV = sd(fev),
          percAsthma = 100*mean(asthma, na.rm = TRUE),
          sdAsthma = sd(asthma, na.rm = TRUE))
```

```
## # A tibble: 2 x 6
##   sex    count meanFEV sdFEV percAsthma sdAsthma
##   <fct>  <int>   <dbl> <dbl>      <dbl>    <dbl>
## 1 male     610   1974.  315.       11.8    0.323
## 2 female   602   2105.  304.       16.4    0.371
```

```
# By obesity level
chsCombo %>%
  group_by(bmiCat) %>%
  summarize(count = n(),
          meanFEV = mean(fev),
          sdFEV = sd(fev),
          percAsthma = 100*mean(asthma, na.rm = TRUE),
          sdAsthma = sd(asthma, na.rm = TRUE))
```

```
## # A tibble: 4 x 6
##   bmiCat      count meanFEV sdFEV percAsthma sdAsthma
##   <fct>       <int>   <dbl> <dbl>      <dbl>    <dbl>
## 1 underweight    35   1699.  305.       8.57    0.284
## 2 normal        987   2011.  295.      13.5     0.342
## 3 overweight     87   2224.  317.      16.1     0.370
## 4 obese         103   2268.  324.      20.4     0.405
```

```
# By smoke and gas exposure
chsCombo %>%
  group_by(smoke_gas_exposure) %>%
  summarize(count = n(),
          meanFEV = mean(fev, na.rm = TRUE),
          sdFEV = sd(fev, na.rm = TRUE),
          percAsthma = 100*mean(asthma, na.rm = TRUE),
          sdAsthma = sd(asthma, na.rm = TRUE))
```

```
## # A tibble: 4 x 6
##   smoke_gas_exposure count meanFEV sdFEV percAsthma sdAsthma
##   <chr>              <int>   <dbl> <dbl>      <dbl>    <dbl>
## 1 both                 154   2034.  301.       12.3    0.330
## 2 gas stove only       803   2031.  318.       14.3    0.351
## 3 neither              219   2066.  328.       14.2    0.349
## 4 smoke only            36   2077.  294.       16.7    0.378
```

Looking at the Data (EDA)

The primary questions of interest are:

1. What is the association between BMI and FEV (forced expiratory volume)?
2. What is the association between smoke and gas exposure and FEV?
3. What is the association between PM2.5 exposure and FEV?

Check Data

```
dim(chsCombo)
```

```
## [1] 1212   52
```

Check Variables

```
str(chsCombo)
```

```
## Classes 'data.table' and 'data.frame':   1212 obs. of  52 variables:
##  $ sid               : int  1 2 6 7 8 10 13 16 19 21 ...
##  $ townname          : chr  "Lancaster" "Lancaster" "Lancaster" "Lancaster" ...
##  $ male              : int  1 1 0 0 0 1 1 0 0 0 ...
##  $ race              : chr  "W" "W" "B" "O" ...
##  $ hispanic          : int  0 0 0 0 1 1 1 0 0 1 ...
##  $ agepft            : num  10.15 10.46 10.1 10.75 9.78 ...
##  $ height            : int  123 145 145 156 132 138 140 141 138 126 ...
##  $ weight            : int  54 77 143 72 61 82 79 74 82 59 ...
##  $ bmi               : num  16.2 16.6 30.9 13.4 15.9 ...
##  $ asthma            : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ active_asthma     : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ father_asthma     : int  0 0 0 NA 1 1 0 0 0 0 ...
##  $ mother_asthma     : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ wheeze            : int  0 1 0 1 1 0 0 1 0 0 ...
##  $ hayfever          : int  0 0 1 0 1 0 0 0 0 0 ...
##  $ allergy           : int  0 0 0 0 1 0 0 1 0 1 ...
##  $ educ_parent       : int  3 5 2 2 3 1 3 3 3 3 ...
##  $ smoke             : int  0 0 0 1 0 0 0 1 0 0 ...
##  $ pets              : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ gasstove          : int  1 0 1 1 0 1 0 1 1 1 ...
##  $ fev               : num  1650 2273 2012 1643 1652 ...
##  $ fvc               : num  1800 2721 2257 2061 1996 ...
##  $ mmef              : num  2538 2366 1819 1462 1607 ...
##  $ pm25_mass         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm25_so4          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm25_no3          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm25_nh4          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm25_oc           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm25_ec           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm25_om           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm10_oc           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm10_ec           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm10_tc           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ formic            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ acetic            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ hcl               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ hno3              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ o3_max            : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
##  $ o3106             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ o3_24             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ no2               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm10              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ no_24hr           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm2_5_fr          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ iacid             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ oacid             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ total_acids       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ lon               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ lat               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ bmiCat            : Factor w/ 4 levels "underweight",..: 2 2 4 1 2 2 2 2 2 2 ...
##  $ smoke_gas_exposure: chr  "gas stove only" "neither" "gas stove only" "both" ...
##  $ sex               : Factor w/ 2 levels "male","female": 2 2 1 1 1 2 2 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(chsCombo)
```

```
##       sid           townname              male            race
##  Min.   :   1.0   Length:1212        Min.   :0.0000   Length:1212
##  1st Qu.: 528.8   Class :character   1st Qu.:0.0000   Class :character
##  Median :1041.5   Mode  :character   Median :0.0000   Mode  :character
##  Mean   :1037.5                      Mean   :0.4967
##  3rd Qu.:1554.2                      3rd Qu.:1.0000
##  Max.   :2053.0                      Max.   :1.0000
##  NA's   :12
##     hispanic          agepft           height          weight
##  Min.   :0.0000   Min.   : 8.961   Min.   :114.0   Min.   : 42.00
##  1st Qu.:0.0000   1st Qu.: 9.634   1st Qu.:135.0   1st Qu.: 66.00
##  Median :0.0000   Median : 9.952   Median :138.0   Median : 76.00
##  Mean   :0.4342   Mean   : 9.927   Mean   :138.9   Mean   : 79.55
##  3rd Qu.:1.0000   3rd Qu.:10.149   3rd Qu.:143.0   3rd Qu.: 87.00
##  Max.   :1.0000   Max.   :12.731   Max.   :165.0   Max.   :207.00
##  NA's   :12
##       bmi            asthma       active_asthma  father_asthma
##  Min.   :11.30   Min.   :0.0000   Min.   :0.00   Min.   :0.00000
##  1st Qu.:15.96   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:0.00000
##  Median :17.85   Median :0.0000   Median :0.00   Median :0.00000
##  Mean   :18.58   Mean   :0.1411   Mean   :0.19   Mean   :0.08318
##  3rd Qu.:19.94   3rd Qu.:0.0000   3rd Qu.:0.00   3rd Qu.:0.00000
##  Max.   :41.27   Max.   :1.0000   Max.   :1.00   Max.   :1.00000
##                                   NA's   :12     NA's   :118
##  mother_asthma        wheeze          hayfever         allergy
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1023   Mean   :0.3313   Mean   :0.1747   Mean   :0.2929
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :68       NA's   :83       NA's   :130      NA's   :75
##   educ_parent        smoke             pets           gasstove
##  Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.0000
##  Median :3.000   Median :0.0000   Median :1.0000   Median :1.0000
##  Mean   :2.747   Mean   :0.1568   Mean   :0.7667   Mean   :0.8267
```

```
##  3rd Qu.:3.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :1.0000   Max.   :1.0000   Max.   :2.0000
##                                   NA's   :12
##       fev              fvc             mmef          pm25_mass
##  Min.   : 984.8   Min.   : 895   Min.   : 757.6   Min.   : 5.960
##  1st Qu.:1829.8   1st Qu.:2067   1st Qu.:2050.8   1st Qu.: 7.615
##  Median :2063.9   Median :2345   Median :2447.5   Median :10.545
##  Mean   :2039.1   Mean   :2335   Mean   :2403.5   Mean   :14.362
##  3rd Qu.:2222.1   3rd Qu.:2547   3rd Qu.:2734.4   3rd Qu.:20.988
##  Max.   :3323.7   Max.   :3698   Max.   :4935.9   Max.   :29.970
##                                                   NA's   :1200
##     pm25_so4        pm25_no3         pm25_nh4         pm25_oc
##  Min.   :0.790   Min.   : 0.730   Min.   :0.4100   Min.   : 1.450
##  1st Qu.:1.077   1st Qu.: 1.538   1st Qu.:0.7375   1st Qu.: 2.520
##  Median :1.815   Median : 2.525   Median :1.1350   Median : 4.035
##  Mean   :1.876   Mean   : 4.488   Mean   :1.7642   Mean   : 4.551
##  3rd Qu.:2.605   3rd Qu.: 7.338   3rd Qu.:2.7725   3rd Qu.: 5.350
##  Max.   :3.230   Max.   :12.200   Max.   :4.2500   Max.   :11.830
##  NA's   :1200    NA's   :1200     NA's   :1200     NA's   :1200
##     pm25_ec          pm25_om          pm10_oc          pm10_ec
##  Min.   :0.1300   Min.   : 1.740   Min.   : 1.860   Min.   :0.1400
##  1st Qu.:0.4000   1st Qu.: 3.020   1st Qu.: 3.228   1st Qu.:0.4100
##  Median :0.5850   Median : 4.840   Median : 5.170   Median :0.5950
##  Mean   :0.7358   Mean   : 5.460   Mean   : 5.832   Mean   :0.7525
##  3rd Qu.:1.1750   3rd Qu.: 6.418   3rd Qu.: 6.855   3rd Qu.:1.1975
##  Max.   :1.3600   Max.   :14.200   Max.   :15.160   Max.   :1.3900
##  NA's   :1200     NA's   :1200     NA's   :1200     NA's   :1200
##     pm10_tc           formic           acetic           hcl
##  Min.   : 1.990   Min.   :0.340   Min.   :0.750   Min.   :0.2200
##  1st Qu.: 3.705   1st Qu.:0.720   1st Qu.:2.297   1st Qu.:0.3250
##  Median : 6.505   Median :1.105   Median :2.910   Median :0.4350
##  Mean   : 6.784   Mean   :1.332   Mean   :3.010   Mean   :0.4208
##  3rd Qu.: 8.430   3rd Qu.:1.765   3rd Qu.:4.000   3rd Qu.:0.4625
##  Max.   :16.440   Max.   :2.770   Max.   :5.140   Max.   :0.7300
##  NA's   :1200     NA's   :1200    NA's   :1200    NA's   :1200
##       hno3           o3_max          o3106           o3_24
##  Min.   :0.430   Min.   :38.27   Min.   :28.22   Min.   :18.22
##  1st Qu.:1.593   1st Qu.:49.93   1st Qu.:41.90   1st Qu.:23.31
##  Median :2.455   Median :64.05   Median :46.74   Median :27.59
##  Mean   :2.367   Mean   :60.16   Mean   :47.76   Mean   :30.23
##  3rd Qu.:3.355   3rd Qu.:67.69   3rd Qu.:55.24   3rd Qu.:32.39
##  Max.   :4.070   Max.   :84.44   Max.   :67.01   Max.   :57.76
##  NA's   :1200    NA's   :1200    NA's   :1200    NA's   :1200
##       no2             pm10          no_24hr          pm2_5_fr
##  Min.   : 4.60   Min.   :18.40   Min.   : 2.050   Min.   : 9.01
##  1st Qu.:12.12   1st Qu.:20.71   1st Qu.: 5.905   1st Qu.:10.28
##  Median :16.40   Median :29.64   Median :12.680   Median :22.23
##  Mean   :18.99   Mean   :32.64   Mean   :16.209   Mean   :19.79
##  3rd Qu.:23.24   3rd Qu.:39.16   3rd Qu.:22.690   3rd Qu.:27.73
##  Max.   :37.97   Max.   :70.39   Max.   :42.950   Max.   :31.55
##  NA's   :1200    NA's   :1200    NA's   :1201     NA's   :1203
##      iacid           oacid        total_acids          lon
##  Min.   :0.760   Min.   :1.090   Min.   : 1.520   Min.   :-120.7
##  1st Qu.:1.835   1st Qu.:2.978   1st Qu.: 4.930   1st Qu.:-118.8
```

```
##  Median :2.825   Median :4.135   Median : 6.370   Median :-117.7
##  Mean   :2.788   Mean   :4.342   Mean   : 6.708   Mean   :-118.3
##  3rd Qu.:3.817   3rd Qu.:5.982   3rd Qu.: 9.395   3rd Qu.:-117.4
##  Max.   :4.620   Max.   :7.400   Max.   :11.430   Max.   :-116.8
##  NA's   :1200    NA's   :1200    NA's   :1200     NA's   :1200
##       lat               bmiCat    smoke_gas_exposure     sex
##  Min.   :32.84   underweight: 35   Length:1212        male  :610
##  1st Qu.:33.93   normal     :987   Class :character   female:602
##  Median :34.10   overweight : 87   Mode  :character
##  Mean   :34.20   obese      :103
##  3rd Qu.:34.65
##  Max.   :35.49
##  NA's   :1200
```

Unlike last assignment, there are no negative minimum values that need to be dealt with. Though there are variables with NAs after the binding of the two data sets, they are within variables that are not used in the analysis of the three questions above. For that, we will keep them as is for now.

Check Variables More Closely

```
# summary() for numerical, table() for categorical
summary(chsCombo$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.30   15.96   17.85   18.58   19.94   41.27
```

```
summary(chsCombo$fev)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   984.8  1829.8  2063.9  2039.1  2222.1  3323.7
```

```
table(chsCombo$smoke_gas_exposure)
```

```
##
##           both gas stove only        neither     smoke only
##            154            803            219             36
```

```
summary(chsCombo$pm25_mass)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   5.960   7.615  10.545  14.362  20.988  29.970    1200
```

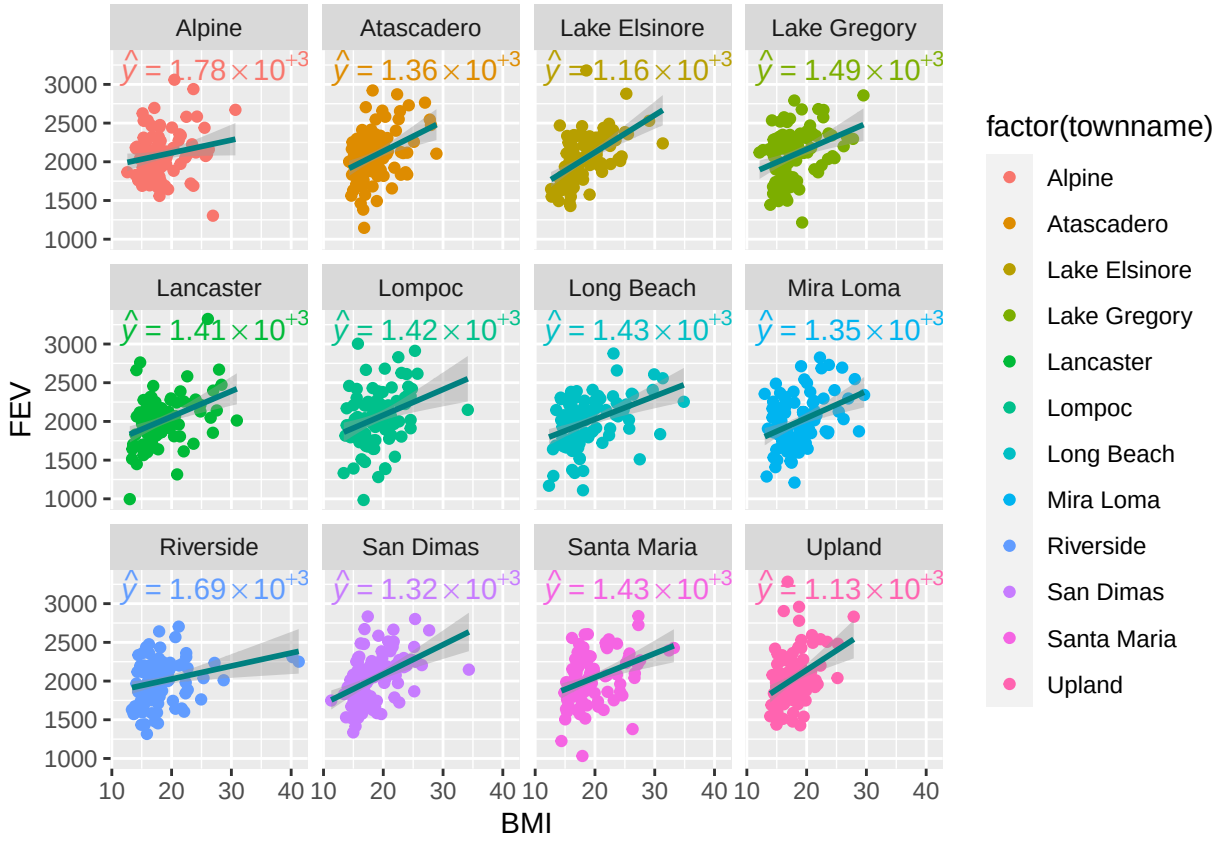
After looking more closely at the variables, each seems adequate with min, max, and other values. However, pm25_mass has 1200 missing values, since the chsInd data set did not provide any data for the variable pm25_mass. We can acknowledge that averaging 12 sites may not be representative to apply to 1200 instances, and since we only are using leaflet() and discovering an association, we can just use the 12 and see how it goes.

Q1

```
chsCombo %>%
  ggplot(mapping = aes(x = bmi, y = fev, color = factor(townname))) +
  geom_point() +
  geom_smooth(method = "lm", color = "#008080") +
  stat_poly_eq(formula = y ~ x,
               eq.with.lhs = "italic(hat(y))~`=`~",
               aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")),
               label.y = 1000,
               parse = TRUE) +
```

```
  labs(x = "BMI", y = "FEV") +
  facet_wrap(~ factor(townname))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

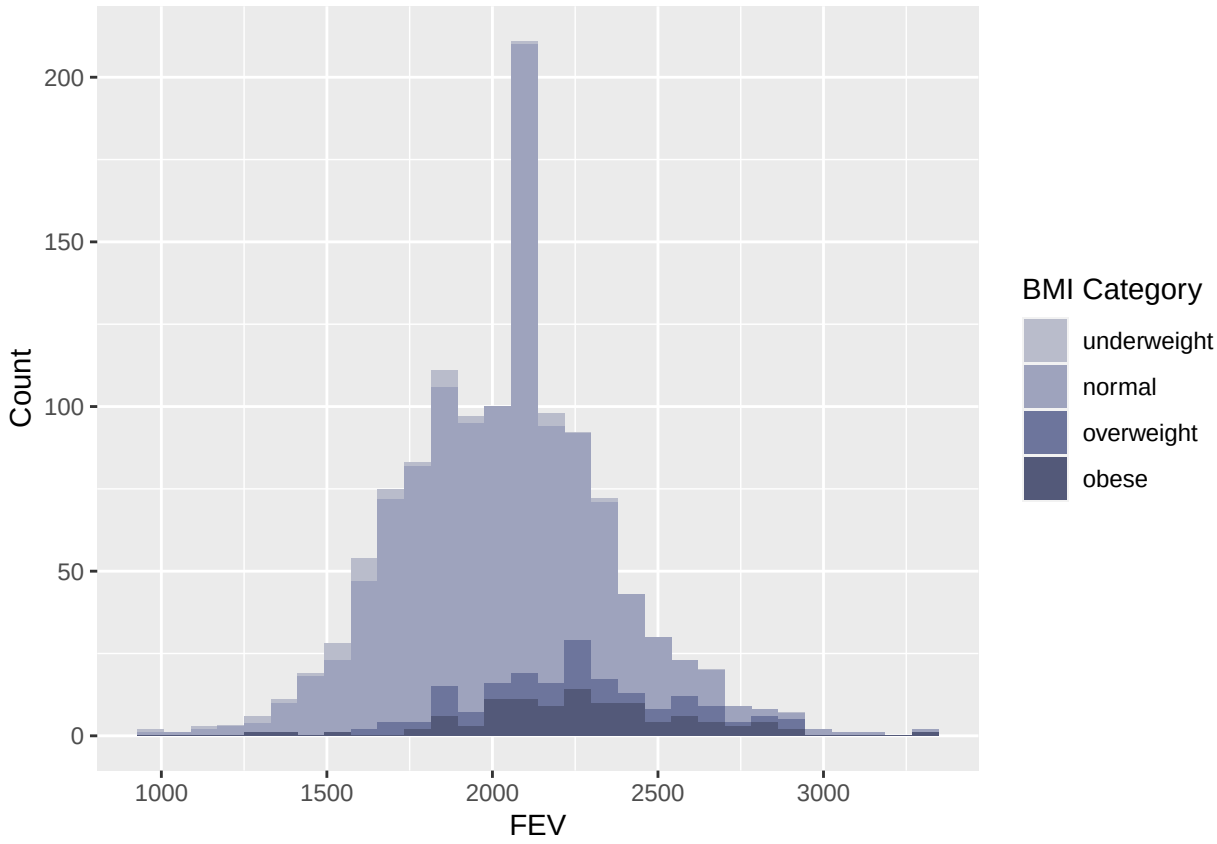


Based on the many scatter plots created, there is a clear relationship that an increase in BMI is associated with an increase in FEV. Of course, some towns experience a stronger association and some weaker. Though the linear association is not visible by points alone, the added linear regression lines appear to add more understanding of the data and the positive relationship involved.

Q2

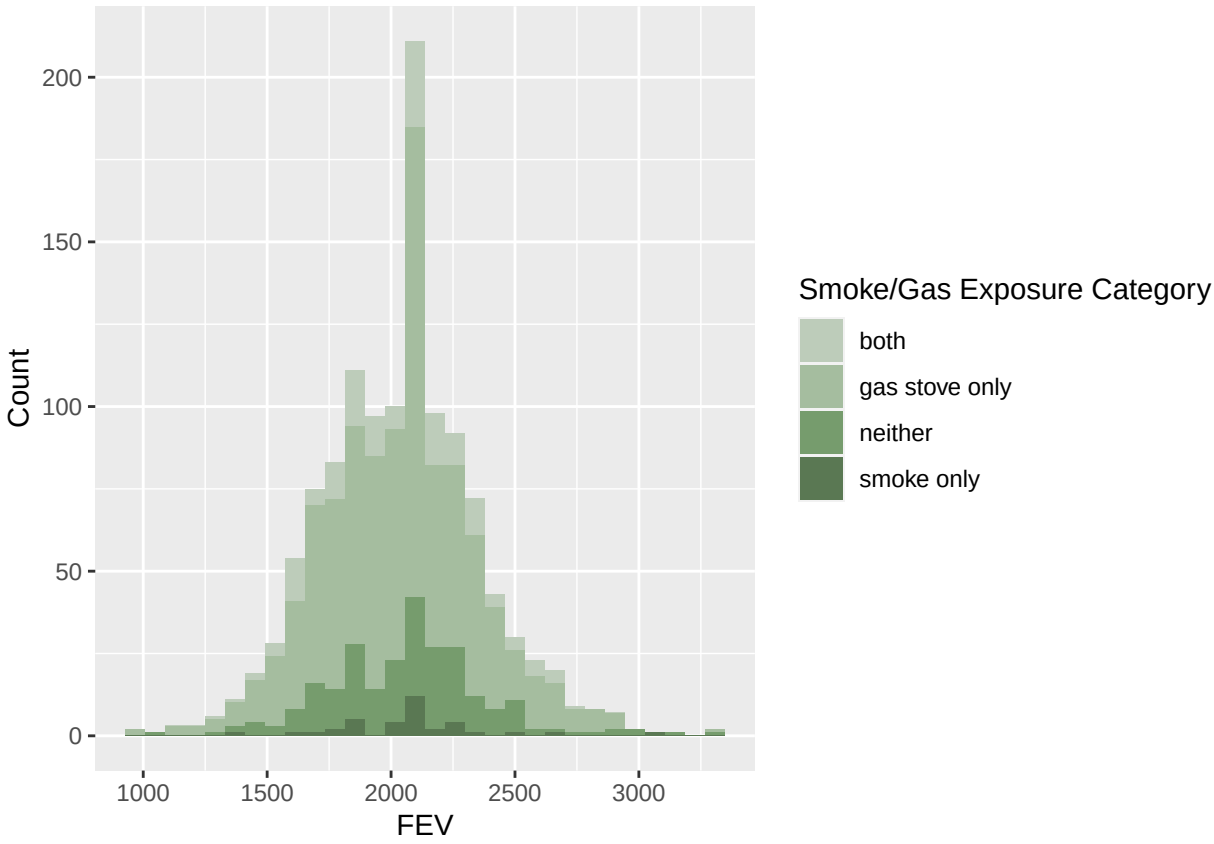
```
chsCombo %>%
  ggplot(aes(x = fev, fill = bmiCat)) +
  geom_histogram() +
  scale_fill_manual(values = c("#B9BCCB","#9EA3BD","#6D759C","#535979")) +
  labs(x = "FEV", y = "Count", fill = "BMI Category")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
chsCombo %>%
  ggplot(aes(x = fev, fill = smoke_gas_exposure)) +
  geom_histogram() +
  scale_fill_manual(values = c("#BDCCBA","#A5BD9F","#769C6D","#5A7853")) +
  labs(x = "FEV", y = "Count", fill = "Smoke/Gas Exposure Category")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

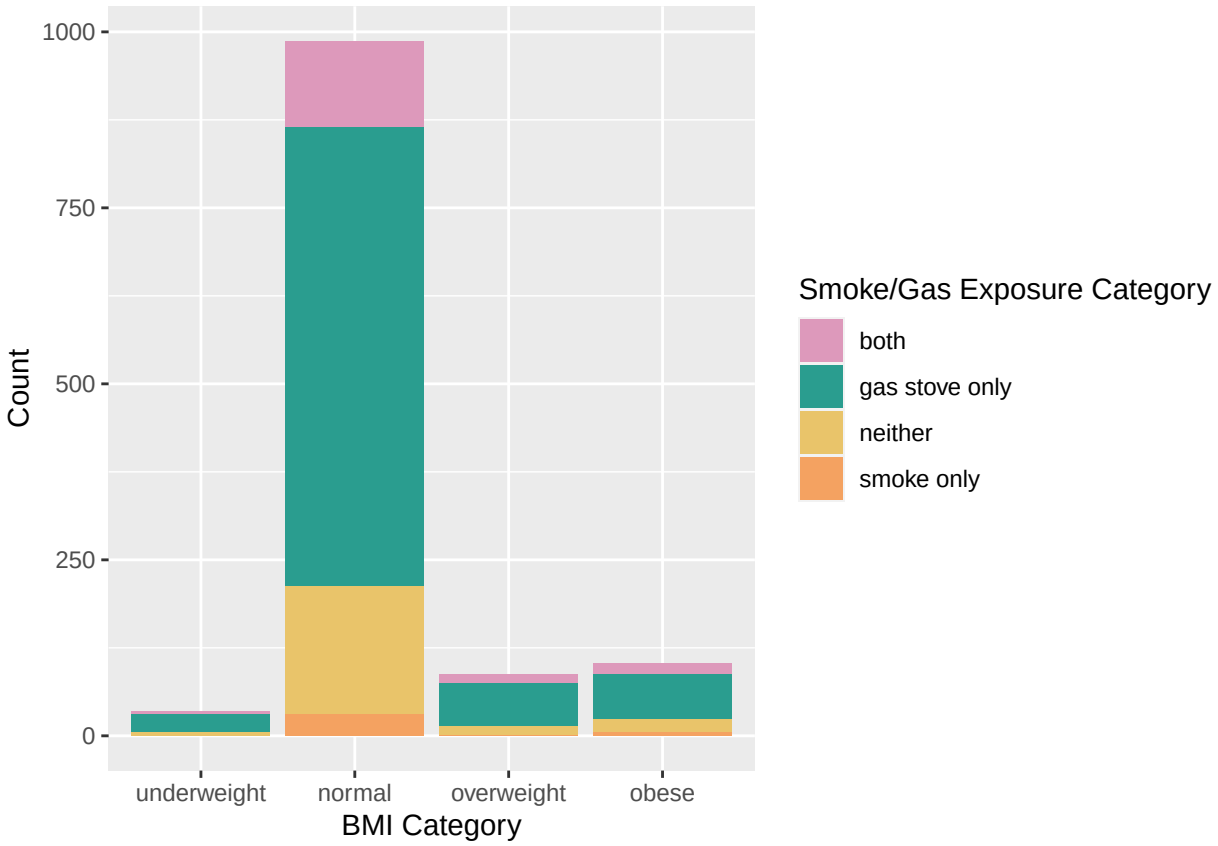


Based on the stacked histogram for BMI category, it is clear that the overweight and obese categories tend to have higher FEV, though they also have the greatest variance across X. Additionally, the plot demonstrates that normal BMI is the most abundant category of the 4. Normal BMI is generally normally distributed for FEV, but it has a clear outlier bin around 2100, which may just be a standard/common value for FEV.

Based on the stacked histogram for smoke/gas exposure category, it appears that each of the four categories have similar means, for no category is as obviously skewed in a direction like that for BMI. The "both" and "gas stove only" categories are the most abundant of the four, and "smoke only" is a rare exposure type.

Q3

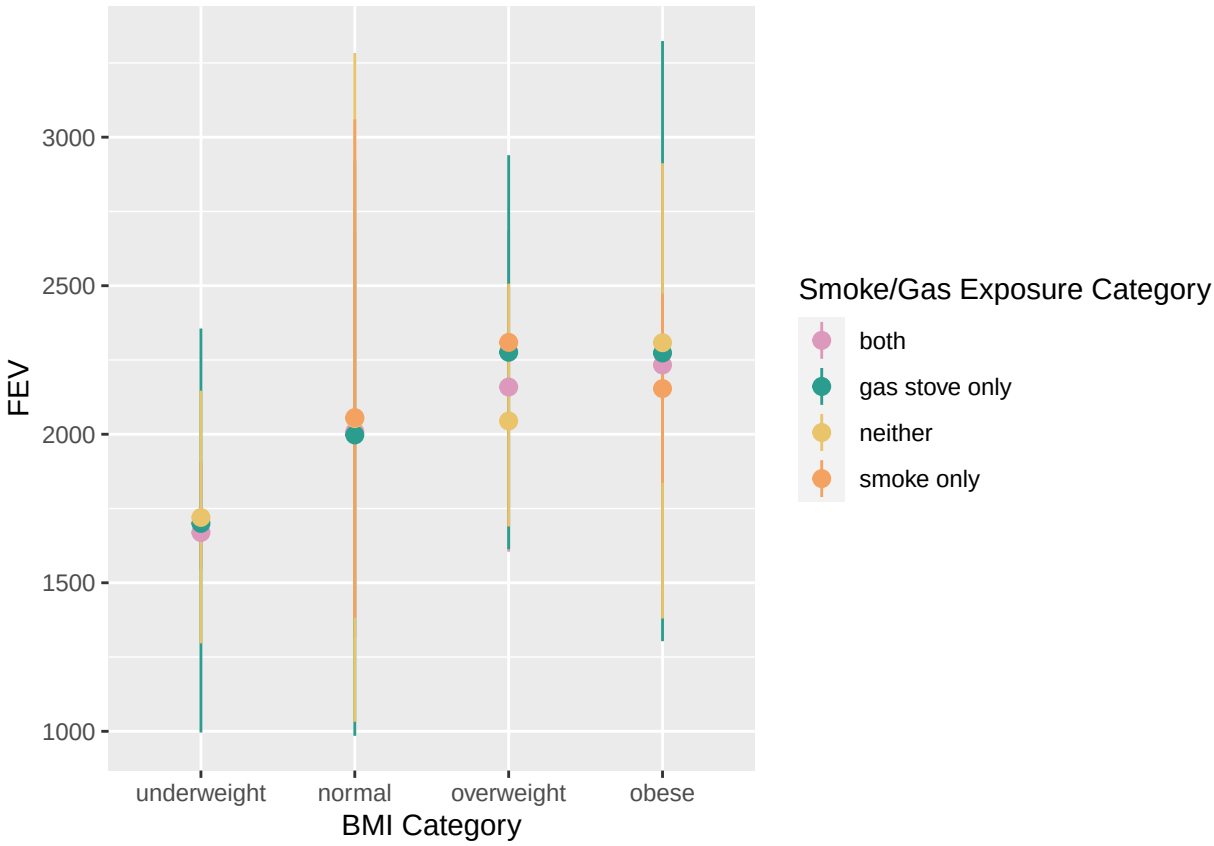
```
chsCombo %>%
  ggplot(aes(x = bmiCat, fill = smoke_gas_exposure)) +
  geom_bar() +
  scale_fill_manual(values = c("#DD99BB","#2A9D8F","#E9C46A","#F4A261")) +
  labs(x = "BMI Category", y = "Count", fill = "Smoke/Gas Exposure Category")
```



Based on the bar graph for BMI category, it is evident that "gas stove only" exposure takes up a similar proportion for each BMI category. Again, "normal" BMI has the greatest number of recorded values of smoke/gas exposure, and therefore may give a more representative idea of how smoke/gas exposure is distributed. The most abundant exposure category is "both," as seen by the pink topping off each created bar.

Q4

```
chsCombo %>%
  ggplot() +
  stat_summary(aes(x = bmiCat, y = fev,
                   group = smoke_gas_exposure,
                   color = smoke_gas_exposure),
                   fun = mean,
                   fun.max = max,
                   fun.min = min) +
  scale_color_manual(values = c("#DD99BB","#2A9D8F","#E9C46A","#F4A261")) +
  labs(x = "BMI Category", y = "FEV", color = "Smoke/Gas Exposure Category")
```

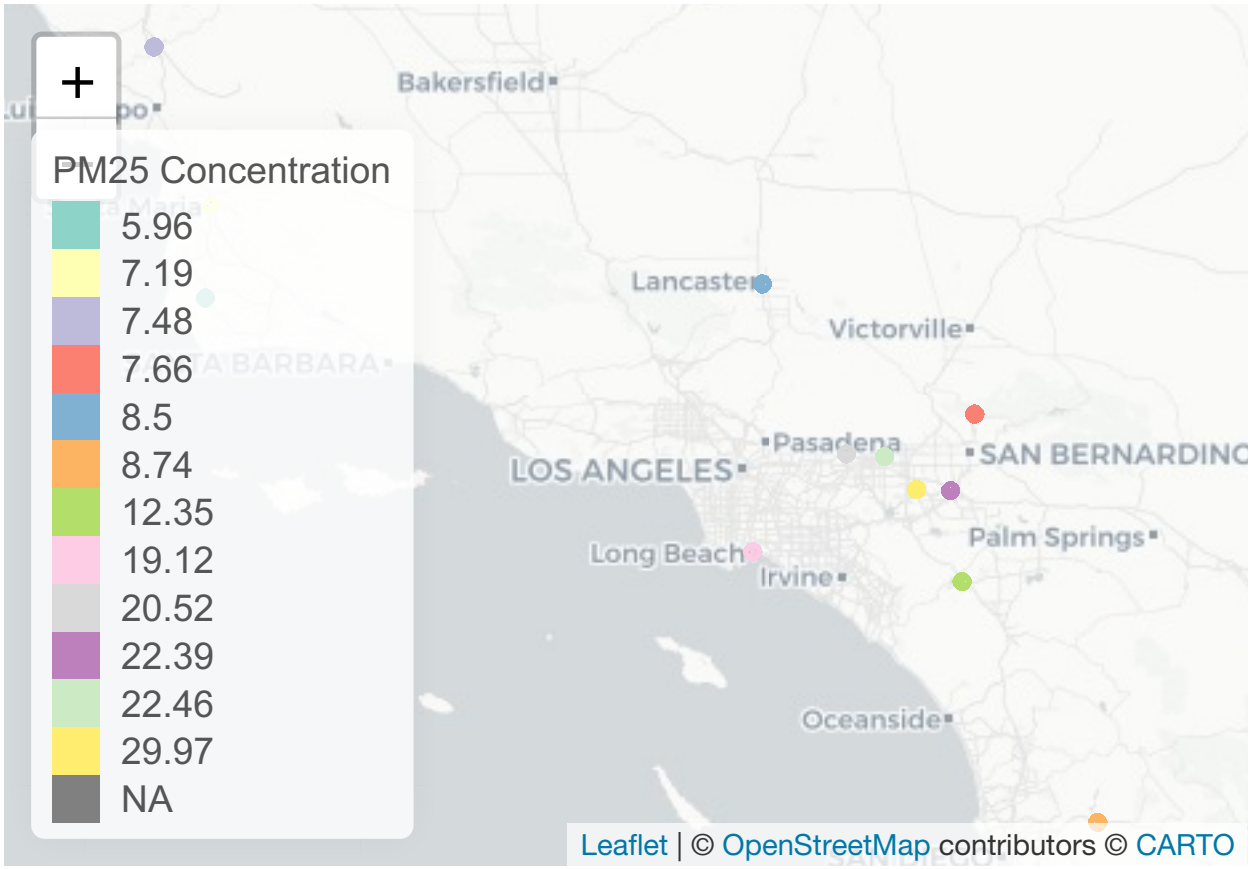


The statistic summary graph for BMI category provides an abundance of information. The lines extending from each plotted average of FEV indicate the minimum and maximum FEV values for each BMI category. In most cases, it appears that the "gas stove only" category experiences the greatest variation in FEV, for its ranges are the largest among the 4 types. The average values of FEV for each category are fairly close to one another, with the "overweight" BMI category demonstrating the greatest differences in means. Furthermore, no particular smoke/gas exposure category appears to be the consistent largest or smallest across BMI category.

Q5

```
temp.pal <- colorFactor(palette ='Set3', domain=chsCombo$pm25_mass)

leaflet(chsCombo[!is.na(pm25_mass)]) %>%
  addProviderTiles('CartoDB.Positron') %>%
  addCircles(
    lat = ~lat, lng=~lon,
    label = chsCombo$townname, color = ~ temp.pal(pm25_mass),
    opacity = .9, fillOpacity = .9, radius = 800) %>%
  addLegend('bottomleft', pal = temp.pal, values = chsCombo$pm25_mass,
          title = 'PM25 Concentration', opacity = 1)
```



This plot gives more insight into geographic location of towns as PM2.5 differs. The lowest PM2.5 values appear closest to Santa Barbara and its neighboring towns. On the other hand, the greatest PM2.5 values appear closest to Los Angeles and the surrounding area.

Q6

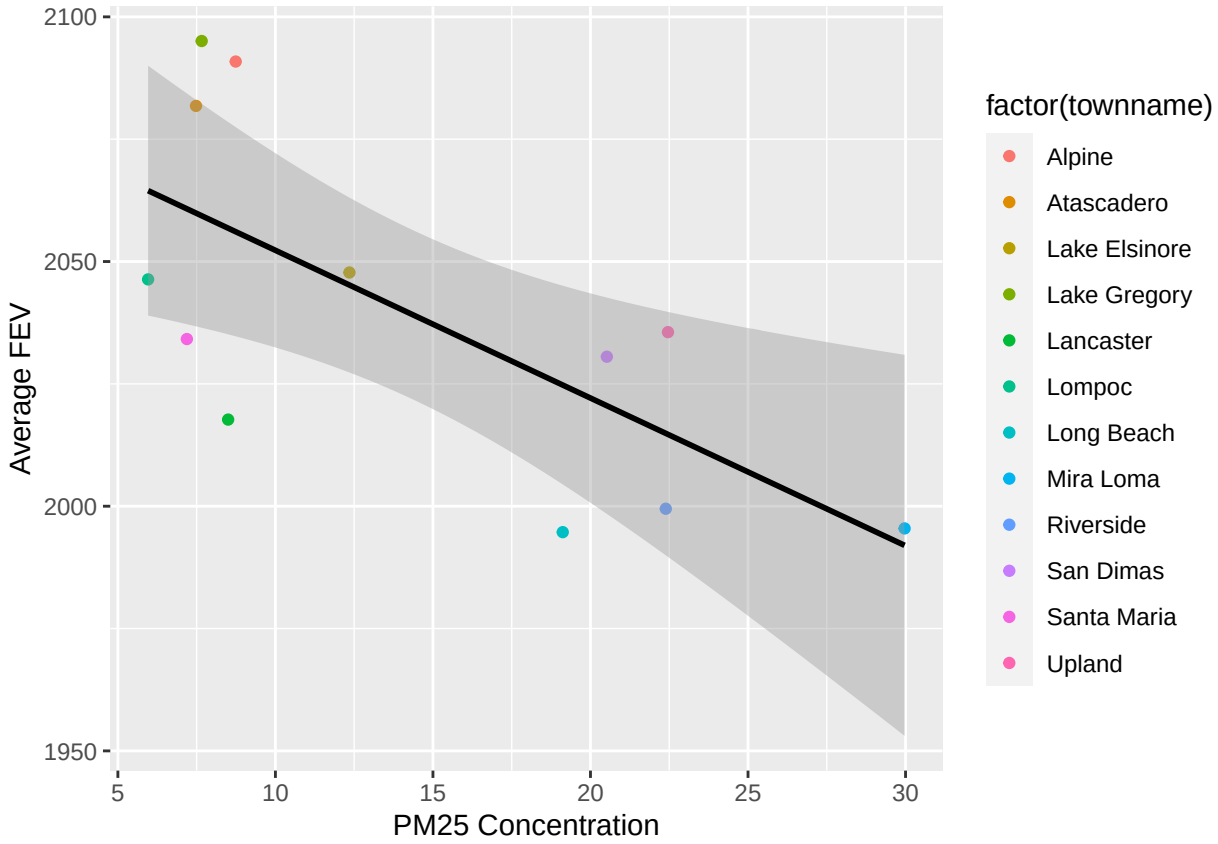
```
chsCombo6 <-
  chsCombo %>%
  group_by(townname) %>%
  mutate(fev = mean(fev))

chsCombo6 %>%
  ggplot(mapping = aes(x = pm25_mass, y = fev, color = factor(townname))) +
  geom_point() +
  geom_smooth(method = "lm", color = "black") +
  labs(x = "PM25 Concentration", y = "Average FEV")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 1200 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1200 rows containing missing values (geom_point).
```



This last plot provides the mean of FEV for each town and each town's PM2.5 concentration. Based on the regression line provided on the plot, it appears that an increase in PM2.5 concentration is associated with a decrease in FEV. While there are certainly larger residuals than others (Lake Gregory, Alpine) and some smaller than others (Mira Loma, Lake Elsinore), the points follow a generally linear pattern. Because of that, it seems evident that the negative correlation exists between the variables FEV and PM2.5.

Overall Question Answers

1. What is the association between BMI and FEV (forced expiratory volume)?

   If someone has a higher BMI, they are more likely to have a higher FEV value.

2. What is the association between smoke and gas exposure and FEV?

   There is less of a clear association between smoke and gas exposure and FEV. However, having "gas stove only" exposure appears to induce a higher FEV than other categories.

3. What is the association between PM2.5 exposure and FEV?

   An increase in PM2.5 exposure is associated with a lower FEV.