

# hw2

AUTHOR  
Andre Gala-Garza

PUBLISHED  
February 14, 2026

**Disclaimer:** Generative AI was used to assist with templating and writing code in this assignment; however, this code was checked manually and edited by hand to ensure accuracy.

**Source:** OpenAI. (2026). *ChatGPT (GPT-5.2 Thinking)* [Large language model]. <https://chatgpt.com/>.

## 1. Data Wrangling

### (i) Merge datasets into one

```
library("dplyr")
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
library("ggplot2")

# load data
chs_individual <- read.csv("chs_individual.csv")
chs_regional   <- read.csv("chs_regional.csv")
```

```
dim(chs_individual)
```

[1] 1200 23

```
head(chs_individual)
```

	sid	townname	male	race	hispanic	agepft	height	weight	bmi	asthma
1	1	Lancaster	1	W	0	10.154689	123	54	16.22411	0
2	2	Lancaster	1	W	0	10.461328	145	77	16.64685	0
3	6	Lancaster	0	B	0	10.097194	145	143	30.91558	0
4	7	Lancaster	0	O	0	10.746064	156	72	13.44809	0
5	8	Lancaster	0	W	1	9.782341	132	61	15.91326	0

6	10	Lancaster	1	0	1	NA	NA	NA	NA	0
		active_asthma	father_asthma	mother_asthma	wheeze	hayfever	allergy	educ_parent		
1		0		0	0	0	0	0		3
2		0		0	0	1	0	0		5
3		0		0	0	0	1	0		2
4		0		NA	0	1	0	0		2
5		0		1	0	1	1	1		3
6		1		1	0	0	0	0		1
		smoke	pets	gasstove	fev	fvc	mmeff			
1	0	1		1	1650.254	1800.005	2537.777			
2	0	1		0	2273.129	2721.111	2365.745			
3	0	0		1	2011.653	2257.244	1818.973			
4	1	1		1	1643.092	2060.526	1462.500			
5	0	1		0	1651.974	1996.382	1606.579			
6	0	1		1	NA	NA	NA			

```
dim(chsRegional)
```

```
[1] 12 27
```

```
head(chsRegional)
```

	townname	pm25_mass	pm25_so4	pm25_no3	pm25_nh4	pm25_oc	pm25_ec	pm25_om			
1	Alpine	8.74	1.73	1.59	0.88	2.54	0.48	3.04			
2	Lake Elsinore	12.35	1.90	2.98	1.36	3.64	0.62	4.36			
3	Lake Gregory	7.66	1.07	2.07	0.91	2.46	0.40	2.96			
4	Lancaster	8.50	0.91	1.87	0.78	4.43	0.55	5.32			
5	Lompoc	5.96	1.08	0.73	0.41	1.45	0.13	1.74			
6	Long Beach	19.12	3.23	6.22	2.57	5.21	1.36	6.25			
	pm10_oc	pm10_ec	pm10_tc	formic	acetic	hcl	hno3	o3_max	o3106	o3_24	no2
1	3.25	0.49	3.75	1.03	2.49	0.41	1.98	65.82	55.05	41.23	12.18
2	4.66	0.63	5.29	1.18	3.56	0.46	2.63	66.70	54.42	32.23	17.03
3	3.16	0.41	3.57	0.66	2.36	0.28	2.28	84.44	67.01	57.76	7.62
4	5.68	0.56	8.61	0.88	2.88	0.22	1.80	54.81	43.88	32.86	15.77
5	1.86	0.14	1.99	0.34	0.75	0.33	0.43	43.85	37.74	28.37	4.60
6	6.68	1.39	8.07	1.57	2.94	0.73	2.67	39.44	28.22	18.22	33.11
	pm10_no_24hr	pm2_5_fr	iacid	oacid	total_acids		lon	lat			
1	24.73	2.48	10.28	2.39	3.52		5.50	-116.7664	32.83505		
2	34.25	7.07	14.53	3.09	4.74		7.37	-117.3273	33.66808		
3	20.05	NA	9.01	2.56	3.02		5.30	-117.2752	34.24290		
4	25.04	12.68	NA	2.02	3.76		5.56	-118.1542	34.68678		
5	18.40	2.05	NA	0.76	1.09		1.52	-120.4579	34.63915		
6	38.41	36.76	22.23	3.40	4.51		7.18	-118.1937	33.77005		

```
names(chsIndividual)
```

```
[1] "sid"                  "townname"             "male"                 "race"
[5] "hispanic"             "agepft"               "height"               "weight"
[9] "bmi"                  "asthma"               "active_asthma"        "father_asthma"
```

```
[13] "mother_asthma" "wheeze"      "hayfever"      "allergy"
[17] "educ_parent"    "smoke"       "pets"         "gasstove"
[21] "fev"           "fvc"        "mmeff"
```

```
names(chsRegional)
```

```
[1] "townname"      "pm25_mass"     "pm25_so4"      "pm25_no3"      "pm25_nh4"
[6] "pm25_oc"       "pm25_ec"       "pm25_om"       "pm10_oc"       "pm10_ec"
[11] "pm10_tc"       "formic"        "acetic"        "hcl"          "hno3"
[16] "o3_max"        "o3106"        "o3_24"        "no2"          "pm10"
[21] "no_24hr"       "pm2_5_fr"     "iacid"        "oacid"        "total_acids"
[26] "lon"           "lat"
```

```
library(dplyr)
```

```
# Check whether townname is unique in regional data
chsRegional %>%
  count(townname, name = "n") %>%
  arrange(desc(n)) %>%
  head()
```

	townname	n
1	Alpine	1
2	Atascadero	1
3	Lake Elsinore	1
4	Lake Gregory	1
5	Lancaster	1
6	Lompoc	1

```
# Merge (adds regional exposures to each child)
chs <- chsIndividual %>%
  left_join(chsRegional, by = "townname")

# Row count should match individual dataset if regional townname is unique
nrow(chsIndividual)
```

```
[1] 1200
```

```
nrow(chs)
```

```
[1] 1200
```

```
# Check for duplicate rows after merge (should be the same as nrow(chs))
nrow(chs)
```

```
[1] 1200
```

```
nrow(distinct(chs))
```

[1] 1200

```
# Check missingness
cat("Rows, cols:\n"); print(dim(chs))
```

Rows, cols:

[1] 1200 49

```
cat("\nFirst 6 rows:\n"); print(head(chs))
```

First 6 rows:

	sid	townname	male	race	hispanic	agepft	height	weight	bmi	asthma	
1	1	Lancaster	1	W	0	10.154689	123	54	16.22411	0	
2	2	Lancaster	1	W	0	10.461328	145	77	16.64685	0	
3	6	Lancaster	0	B	0	10.097194	145	143	30.91558	0	
4	7	Lancaster	0	O	0	10.746064	156	72	13.44809	0	
5	8	Lancaster	0	W	1	9.782341	132	61	15.91326	0	
6	10	Lancaster	1	O	1	NA	NA	NA	NA	0	
	active_asthma	father_asthma	mother_asthma	wheeze	hayfever	allergy	educ_parent				
1	0	0	0	0	0	0	0			3	
2	0	0	0	0	1	0	0			5	
3	0	0	0	0	0	1	0			2	
4	0		NA	0	1	0	0			2	
5	0		1	0	1	1	1			3	
6	1		1	0	0	0	0			1	
	smoke	pets	gasstove	fev	fvc	mmeff	pm25_mass	pm25_so4	pm25_no3		
1	0	1	1	1650.254	1800.005	2537.777	8.5	0.91	1.87		
2	0	1	0	2273.129	2721.111	2365.745	8.5	0.91	1.87		
3	0	0	1	2011.653	2257.244	1818.973	8.5	0.91	1.87		
4	1	1	1	1643.092	2060.526	1462.500	8.5	0.91	1.87		
5	0	1	0	1651.974	1996.382	1606.579	8.5	0.91	1.87		
6	0	1	1	NA	NA	NA	8.5	0.91	1.87		
	pm25_nh4	pm25_oc	pm25_ec	pm25_om	pm10_oc	pm10_ec	pm10_tc	formic	acetic	hcl	
1	0.78	4.43	0.55	5.32	5.68	0.56	8.61	0.88	2.88	0.22	
2	0.78	4.43	0.55	5.32	5.68	0.56	8.61	0.88	2.88	0.22	
3	0.78	4.43	0.55	5.32	5.68	0.56	8.61	0.88	2.88	0.22	
4	0.78	4.43	0.55	5.32	5.68	0.56	8.61	0.88	2.88	0.22	
5	0.78	4.43	0.55	5.32	5.68	0.56	8.61	0.88	2.88	0.22	
6	0.78	4.43	0.55	5.32	5.68	0.56	8.61	0.88	2.88	0.22	
	hno3	o3_max	o3106	o3_24	no2	pm10	no_24hr	pm2_5_fr	iacid	oacid	total_acids
1	1.8	54.81	43.88	32.86	15.77	25.04	12.68	NA	2.02	3.76	5.56
2	1.8	54.81	43.88	32.86	15.77	25.04	12.68	NA	2.02	3.76	5.56
3	1.8	54.81	43.88	32.86	15.77	25.04	12.68	NA	2.02	3.76	5.56
4	1.8	54.81	43.88	32.86	15.77	25.04	12.68	NA	2.02	3.76	5.56

```

5 1.8 54.81 43.88 32.86 15.77 25.04 12.68      NA 2.02 3.76      5.56
6 1.8 54.81 43.88 32.86 15.77 25.04 12.68      NA 2.02 3.76      5.56
    lon      lat
1 -118.1542 34.68678
2 -118.1542 34.68678
3 -118.1542 34.68678
4 -118.1542 34.68678
5 -118.1542 34.68678
6 -118.1542 34.68678

```

```

na_counts <- sapply(chs, function(x) sum(is.na(x)))
na_counts <- sort(na_counts[na_counts > 0], decreasing = TRUE)

cat("\nVariables with missing values (count):\n")

```

Variables with missing values (count):

```
print(na_counts)
```

pm2_5_fr	hayfever	father_asthma	mmef	no_24hr
300	118	106	106	100
fvc	fev	agepft	height	weight
97	95	89	89	89
bmi	wheeze	educ_parent	allergy	mother_asthma
89	71	64	63	56
smoke	gasstove	asthma		
40	33	31		

```
cat("\nTotal missing values:\n")
```

Total missing values:

```
print(sum(is.na(chs)))
```

[1] 1636

```

# ---- imputation helpers ----
Mode <- function(x) {
  x <- x[!is.na(x)]
  if (length(x) == 0) return(NA)
  tab <- table(x)
  names(tab)[which.max(tab)]
}

# Identify variable types
num_vars <- names(chs)[sapply(chs, is.numeric)]

```

```
cat_vars <- names(chs)[sapply(chs, function(x) is.character(x) || is.factor(x) || is.logical(x))]

# Don't try to impute the grouping variables with themselves
num_vars <- setdiff(num_vars, c("male", "hispanic"))
cat_vars <- setdiff(cat_vars, c("male", "hispanic"))

cat("\nNumeric variables to mean-impute within (male,hispanic):\n")
```

Numeric variables to mean-impute within (male,hispanic):

```
print(num_vars)
```

```
[1] "sid"          "agepft"       "height"        "weight"
[5] "bmi"          "asthma"        "active_asthma" "father_asthma"
[9] "mother_asthma" "wheeze"        "hayfever"      "allergy"
[13] "educ_parent"   "smoke"         "pets"          "gasstove"
[17] "fev"           "fvc"           "mmef"          "pm25_mass"
[21] "pm25_so4"     "pm25_no3"      "pm25_nh4"      "pm25_oc"
[25] "pm25_ec"       "pm25_om"        "pm10_oc"       "pm10_ec"
[29] "pm10_tc"       "formic"        "acetic"        "hcl"
[33] "hno3"          "o3_max"        "o3106"        "o3_24"
[37] "no2"           "pm10"          "no_24hr"       "pm2_5_fr"
[41] "iacid"         "oacid"         "total_acids"  "lon"
[45] "lat"
```

```
cat("\nCategorical variables to mode-impute within (male,hispanic):\n")
```

Categorical variables to mode-impute within (male,hispanic):

```
print(cat_vars)
```

```
[1] "townname" "race"
```

```
# Groupwise imputation: mean for numeric, mode for categorical
chs_imp <- chs %>%
  group_by(male, hispanic) %>%
  mutate(
    across(all_of(num_vars),
      ~ ifelse(is.na(.x), mean(.x, na.rm = TRUE), .x)),
    across(all_of(cat_vars),
      ~ {
        m <- Mode(.x)
        ifelse(is.na(.x), m, .x)
      })
  ) %>%
  ungroup()
```

```
# if some groups had all-NA for a variable, mean(.x, na.rm=TRUE) returns NaN
# and Mode() returns NA. Replace any remaining NaN with NA to make it obvious.
chs_imp <- chs_imp %>%
  mutate(across(where(is.numeric), ~ ifelse(is.nan(.x), NA, .x)))

# confirm missingness after imputation
na_counts_after <- sapply(chs_imp, function(x) sum(is.na(x)))
na_counts_after <- sort(na_counts_after[na_counts_after > 0], decreasing = TRUE)

cat("\nRemaining variables with missing values after imputation (count):\n")
```

Remaining variables with missing values after imputation (count):

```
print(na_counts_after)

named integer(0)

cat("\nTotal missing values after imputation:\n")
```

Total missing values after imputation:

```
print(sum(is.na(chs_imp)))
```

[1] 0

```
cat("\nHead of imputed dataset:\n")
```

Head of imputed dataset:

```
head(chs_imp)

# A tibble: 6 × 49
  sid townname male race hispanic agepft height weight bmi asthma
  <int> <chr>    <int> <chr>    <int> <dbl>   <dbl> <dbl> <dbl> <dbl>
1     1 Lancaster     1 W          0 10.2    123    54  16.2     0
2     2 Lancaster     1 W          0 10.5    145    77  16.6     0
3     6 Lancaster     0 B          0 10.1    145   143  30.9     0
4     7 Lancaster     0 O          0 10.7    156    72  13.4     0
5     8 Lancaster     0 W          1  9.78   132    61  15.9     0
6    10 Lancaster     1 O          1  9.97   139.   82.8 19.4     0
# i 39 more variables: active_asthma <int>, father_asthma <dbl>,
#   mother_asthma <dbl>, wheeze <dbl>, hayfever <dbl>, allergy <dbl>,
#   educ_parent <dbl>, smoke <dbl>, pets <int>, gasstove <dbl>, fev <dbl>,
#   fvc <dbl>, mmeff <dbl>, pm25_mass <dbl>, pm25_so4 <dbl>, pm25_no3 <dbl>,
```

```
# pm25_nh4 <dbl>, pm25_oc <dbl>, pm25_ec <dbl>, pm25_om <dbl>, pm10_oc <dbl>,
# pm10_ec <dbl>, pm10_tc <dbl>, formic <dbl>, acetic <dbl>, hcl <dbl>,
# hno3 <dbl>, o3_max <dbl>, o3106 <dbl>, o3_24 <dbl>, no2 <dbl>, ...
```

## (ii) Create obesity level categorical variable

```
names(chs_imp)
```

```
[1] "sid"          "townname"      "male"         "race"
[5] "hispanic"     "agepft"        "height"       "weight"
[9] "bmi"          "asthma"        "active_asthma" "father_asthma"
[13] "mother_asthma" "wheeze"        "hayfever"     "allergy"
[17] "educ_parent"   "smoke"         "pets"         "gasstove"
[21] "fev"           "fvc"           "mmef"         "pm25_mass"
[25] "pm25_so4"      "pm25_no3"      "pm25_nh4"     "pm25_oc"
[29] "pm25_ec"       "pm25_om"        "pm10_oc"      "pm10_ec"
[33] "pm10_tc"       "formic"        "acetic"       "hcl"
[37] "hno3"          "o3_max"        "o3106"       "o3_24"
[41] "no2"           "pm10"          "no_24hr"     "pm2_5_fr"
[45] "iacid"         "oacid"         "total_acids" "lon"
[49] "lat"
```

```
bmi_var <- "bmi"

chs_imp <- chs_imp %>%
  mutate(
    obesity_level = case_when(
      .data[[bmi_var]] < 14 ~ "underweight",
      .data[[bmi_var]] >= 14 & .data[[bmi_var]] <= 22 ~ "normal",
      .data[[bmi_var]] > 22 & .data[[bmi_var]] <= 24 ~ "overweight",
      .data[[bmi_var]] > 24 ~ "obese",
      TRUE ~ NA_character_
    ),
    obesity_level = factor(obesity_level,
                           levels = c("underweight", "normal", "overweight", "obese"))
  )

head(chs_imp %>% select(all_of(bmi_var), obesity_level))
```

```
# A tibble: 6 × 2
  bmi obesity_level
  <dbl> <fct>
1 16.2 normal
2 16.6 normal
3 30.9 obese
4 13.4 underweight
5 15.9 normal
6 19.4 normal
```

```
# Summary table to verify coding
obesity_check <- chs_imp %>%
  group_by(obesity_level) %>%
  summarise(
    min_BMI = min(.data[[bmi_var]], na.rm = TRUE),
    max_BMI = max(.data[[bmi_var]], na.rm = TRUE),
    n = n(),
    .groups = "drop"
  )
```

```
obesity_check
```

```
# A tibble: 4 × 4
  obesity_level min_BMI max_BMI     n
  <fct>        <dbl>   <dbl> <int>
1 underweight    11.3    14.0    35
2 normal         14.0    22.0   975
3 overweight     22.0    24.0    87
4 obese          24.0    41.3   103
```

### (iii) Create smoke/gas exposure categorical variable

```
library(dplyr)

names(chs_imp)
```

```
[1] "sid"           "townname"       "male"           "race"
[5] "hispanic"      "agepft"         "height"         "weight"
[9] "bmi"            "asthma"         "active_asthma" "father_asthma"
[13] "mother_asthma" "wheeze"          "hayfever"       "allergy"
[17] "educ_parent"   "smoke"          "pets"           "gasstove"
[21] "fev"            "fvc"             "mmef"          "pm25_mass"
[25] "pm25_so4"       "pm25_no3"        "pm25_nh4"      "pm25_oc"
[29] "pm25_ec"        "pm25_om"         "pm10_oc"        "pm10_ec"
[33] "pm10_tc"        "formic"         "acetic"         "hcl"
[37] "hno3"           "o3_max"         "o3106"         "o3_24"
[41] "no2"            "pm10"           "no_24hr"       "pm2_5_fr"
[45] "iacid"          "oacid"          "total_acids"  "lon"
[49] "lat"            "obesity_level"
```

```
shs_var <- "smoke"
gas_var <- "gasstove"

chs_imp <- chs_imp %>%
  mutate(
    shs_bin = case_when(
      is.na(.data[[shs_var]]) ~ NA,
      .data[[shs_var]] %in% c(1, "1", "Y", "Yes", "YES", TRUE) ~ 1,
```

```
.data[[shs_var]] %in% c(0, "0", "N", "No", "NO", FALSE) ~ 0,
  TRUE ~ NA_real_
),
gas_bin = case_when(
  is.na(.data[[gas_var]]) ~ NA,
  .data[[gas_var]] %in% c(1, "1", "Y", "Yes", "YES", TRUE) ~ 1,
  .data[[gas_var]] %in% c(0, "0", "N", "No", "NO", FALSE) ~ 0,
  TRUE ~ NA_real_
),
smoke_gas_exposure = case_when(
  is.na(shs_bin) | is.na(gas_bin) ~ NA_character_,
  shs_bin == 0 & gas_bin == 0 ~ "neither",
  shs_bin == 1 & gas_bin == 0 ~ "secondhand_smoke_only",
  shs_bin == 0 & gas_bin == 1 ~ "gas_stove_only",
  shs_bin == 1 & gas_bin == 1 ~ "both",
  TRUE ~ NA_character_
),
smoke_gas_exposure = factor(
  smoke_gas_exposure,
  levels = c("neither", "secondhand_smoke_only", "gas_stove_only", "both")
)
) %>%
  select(-shs_bin, -gas_bin)

head(chs_imp %>% select(all_of(shs_var), all_of(gas_var), smoke_gas_exposure))
```

```
# A tibble: 6 × 3
  smoke gasstove smoke_gas_exposure
  <dbl>    <dbl> <fct>
1     0        1 gas_stove_only
2     0        0 neither
3     0        1 gas_stove_only
4     1        1 both
5     0        0 neither
6     0        1 gas_stove_only
```

```
# Quick check: counts in the four categories
chs_imp %>%
  count(smoke_gas_exposure, name = "n")
```

```
# A tibble: 5 × 2
  smoke_gas_exposure      n
  <fct>                  <int>
1 neither                  214
2 secondhand_smoke_only    36
3 gas_stove_only           739
4 both                      151
5 <NA>                      60
```

```
unique(chs_imp$gasstove)
```

```
[1] 1.0000000 0.0000000 0.8156863 0.8218623 0.7798742 0.7291066
```

```
unique(chs_imp$smoke)
```

```
[1] 0.0000000 1.0000000 0.1535270 0.1949686 0.1501976 0.1522989
```

Note that for some of the records in the dataset, the values of "gasstove" and "smoke" were not binary (i.e. either 0 or 1). For these cases, the "smoke\_gas\_exposure" categorical variable was set to NA.

## (iv) Make four summary tables

```
library(dplyr)
```

```
# Asthma indicator: Forced expiratory volume in 1 second (ml)
names(chs_imp)
```

```
[1] "sid"                  "townname"            "male"
[4] "race"                 "hispanic"            "agepft"
[7] "height"               "weight"              "bmi"
[10] "asthma"               "active_asthma"       "father_asthma"
[13] "mother_asthma"        "wheeze"              "hayfever"
[16] "allergy"              "educ_parent"         "smoke"
[19] "pets"                 "gasstove"            "fev"
[22] "fvc"                  "mmef"                "pm25_mass"
[25] "pm25_so4"             "pm25_no3"            "pm25_nh4"
[28] "pm25_oc"               "pm25_ec"              "pm25_om"
[31] "pm10_oc"               "pm10_ec"              "pm10_tc"
[34] "formic"                "acetic"              "hcl"
[37] "hno3"                  "o3_max"              "o3106"
[40] "o3_24"                 "no2"                 "pm10"
[43] "no_24hr"               "pm2_5_fr"            "iacid"
[46] "oacid"                 "total_acids"         "lon"
[49] "lat"                   "obesity_level"       "smoke_gas_exposure"
```

```
summary(chs_imp$fev)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
984.8	1827.6	2016.4	2030.1	2223.6	3323.7

```
# Make sex readable (male is 0/1)
chs_imp <- chs_imp %>%
  mutate(sex = factor(male, levels = c(0, 1), labels = c("female", "male")))

# Helper for mean/sd table
fev_summary <- function(df, group_var) {
```

```

df %>%
  group_by(across(all_of(group_var))) %>%
  summarise(
    mean_fev = mean(fev, na.rm = TRUE),
    sd_fev   = sd(fev, na.rm = TRUE),
    n         = sum(!is.na(fev)),
    .groups   = "drop"
  ) %>%
  arrange(mean_fev)
}

# 1) By town
tab_town <- fev_summary(chs_imp, "townname")
tab_town

```

# A tibble: 12 × 4

	townname	mean_fev	sd_fev	n
	<chr>	<dbl>	<dbl>	<int>
1	Mira Loma	1985.	325.	100
2	Long Beach	1986.	319.	100
3	Riverside	1990.	278.	100
4	Lancaster	2003.	317.	100
5	Upland	2024.	343.	100
6	Santa Maria	2026.	312.	100
7	San Dimas	2027.	319.	100
8	Lompoc	2034.	351.	100
9	Lake Elsinore	2039.	304.	100
10	Atascadero	2076.	324.	100
11	Lake Gregory	2085.	320.	100
12	Alpine	2087.	291.	100

```

# 2) By sex
tab_sex <- fev_summary(chs_imp, "sex")
tab_sex

```

# A tibble: 2 × 4

	sex	mean_fev	sd_fev	n
	<fct>	<dbl>	<dbl>	<int>
1	female	1959.	312.	610
2	male	2104.	308.	590

```

# 3) By obesity level
tab_obesity <- fev_summary(chs_imp, "obesity_level")
tab_obesity

```

# A tibble: 4 × 4

	obesity_level	mean_fev	sd_fev	n
	<fct>	<dbl>	<dbl>	<int>
1	underweight	1698.	303.	35

2 normal	2000.	295.	975
3 overweight	2224.	317.	87
4 obese	2266.	325.	103

```
# 4) By smoke_gas_exposure
tab_smoke_gas <- fev_summary(chs_imp, "smoke_gas_exposure")
tab_smoke_gas
```

```
# A tibble: 5 × 4
  smoke_gas_exposure    mean_fev   sd_fev     n
  <fct>                 <dbl>     <dbl>   <int>
1 <NA>                  2002.     340.     60
2 both                   2020.     299.    151
3 gas_stove_only         2026.     318.    739
4 neither                2055.     330.    214
5 secondhand_smoke_only  2056.     296.     36
```

In the town table, we see that the mean FEV varies from 1985 to 2075 ml, but the standard distribution is often close to or slightly smaller than that off the sex and obesity tables, so the town alone is not explaining most of the variability in FEV.

In the sex table, we see that males have a noticeably higher FEV than females (2103 ml compared to 1958 ml), while the standard deviations are similar.

In the smoke/gas exposure table, we see that the means are very similar across groups relative to the standard deviation. Combined with the fact that some groups have very small sample sizes, this means that the group means can bounce around a lot just due to random sampling.

## 2. Exploratory Data Analysis (EDA)

### Steps 1-4

---

From the previous problem, we have already completed the first four steps on the EDA checklist: 1. Read in the data 2. Check the size of the data 3. Examine the variables and their types 4. Look at the top and bottom of the data

### Step 5: Visualize distributions of key variables

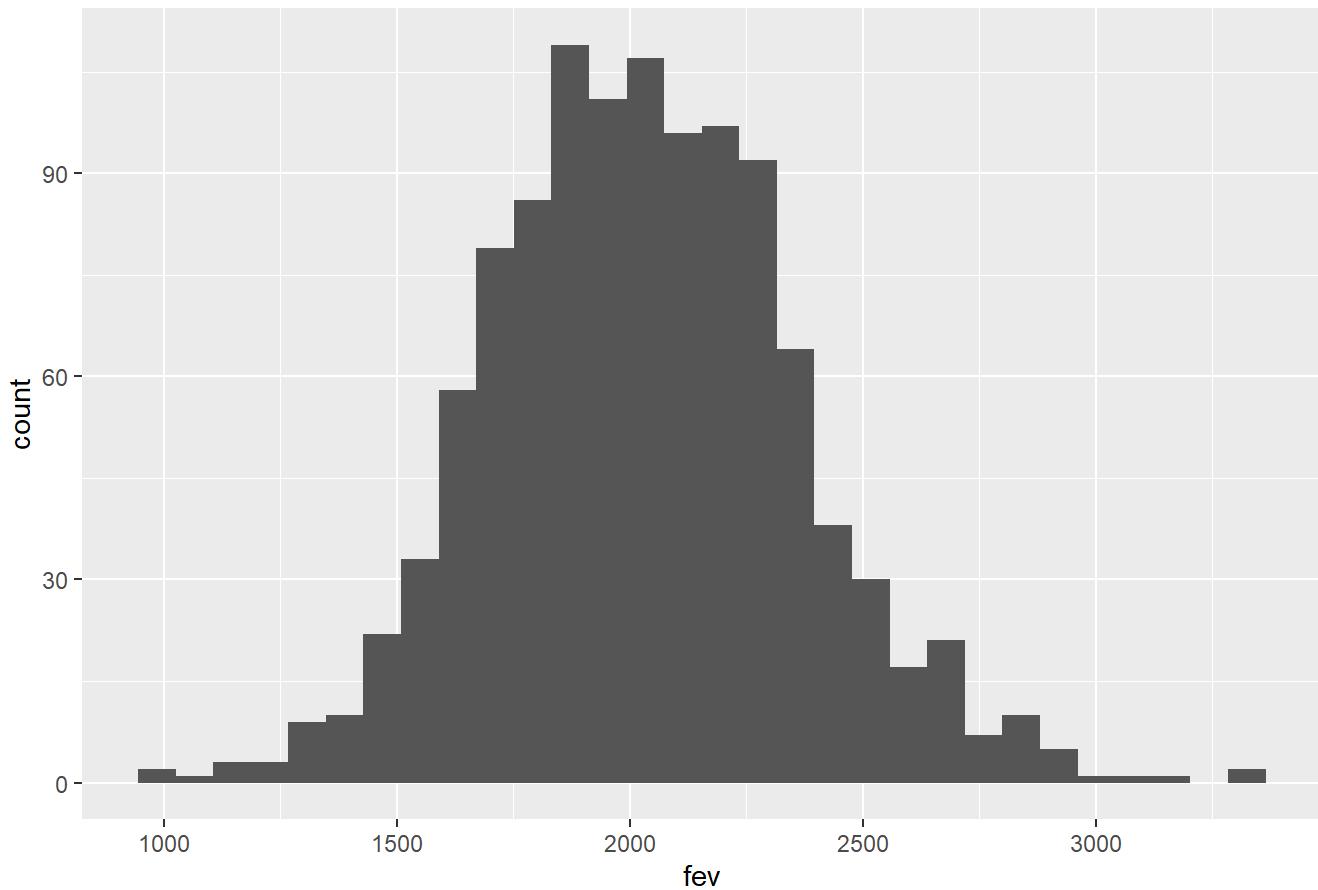
---

```
# FEV
ggplot(chs, aes(x = fev)) + geom_histogram() + labs(title = "Distribution of FEV")
```

`stat\_bin()` using `bins = 30`. Pick better value `binwidth`.

Warning: Removed 95 rows containing non-finite outside the scale range  
(`stat\_bin()`).

## Distribution of FEV

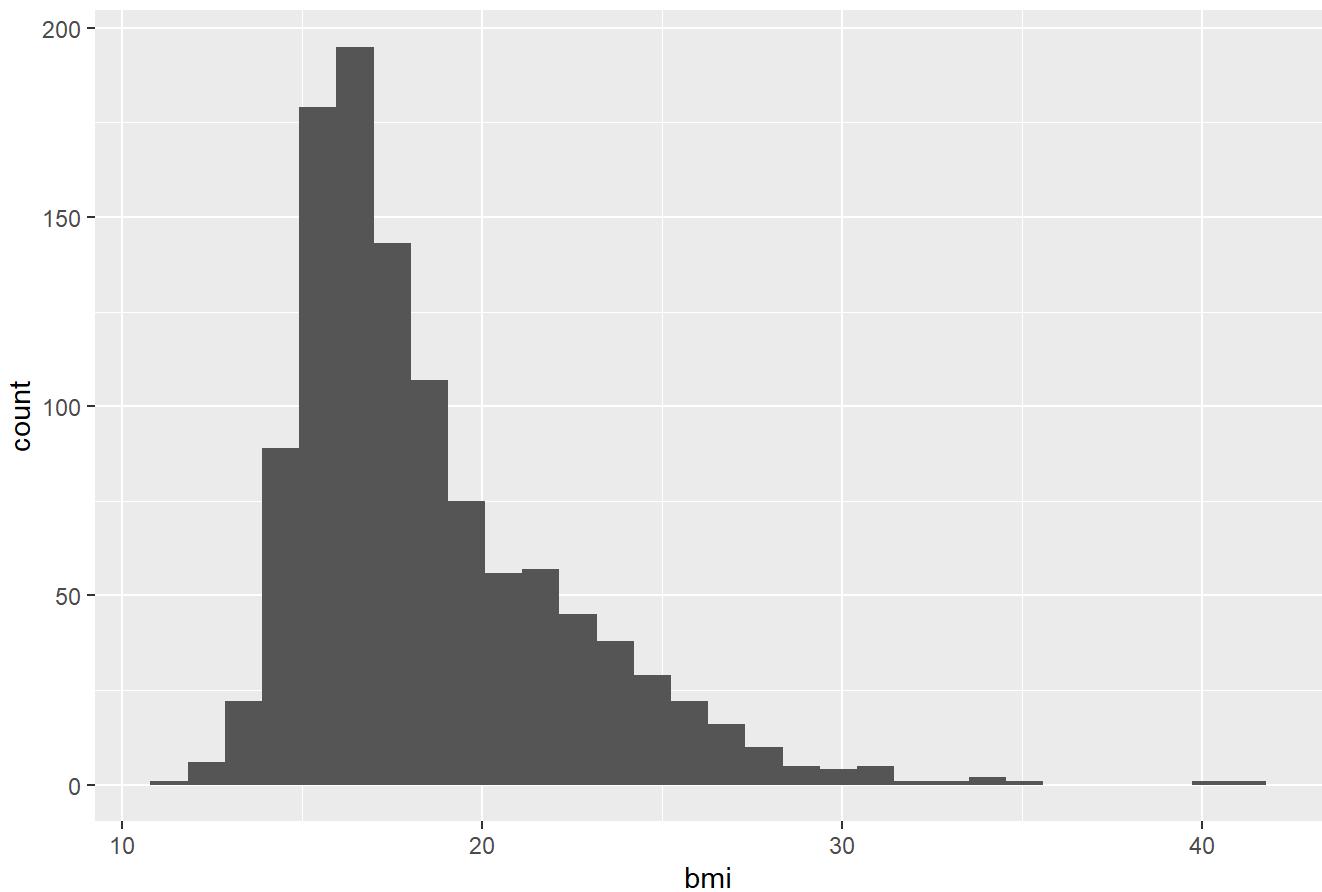


```
# BMI  
ggplot(chs, aes(x = bmi)) + geom_histogram() + labs(title = "Distribution of BMI")
```

`stat\_bin()` using `bins = 30`. Pick better value `binwidth`.

Warning: Removed 89 rows containing non-finite outside the scale range  
(`stat\_bin()`).

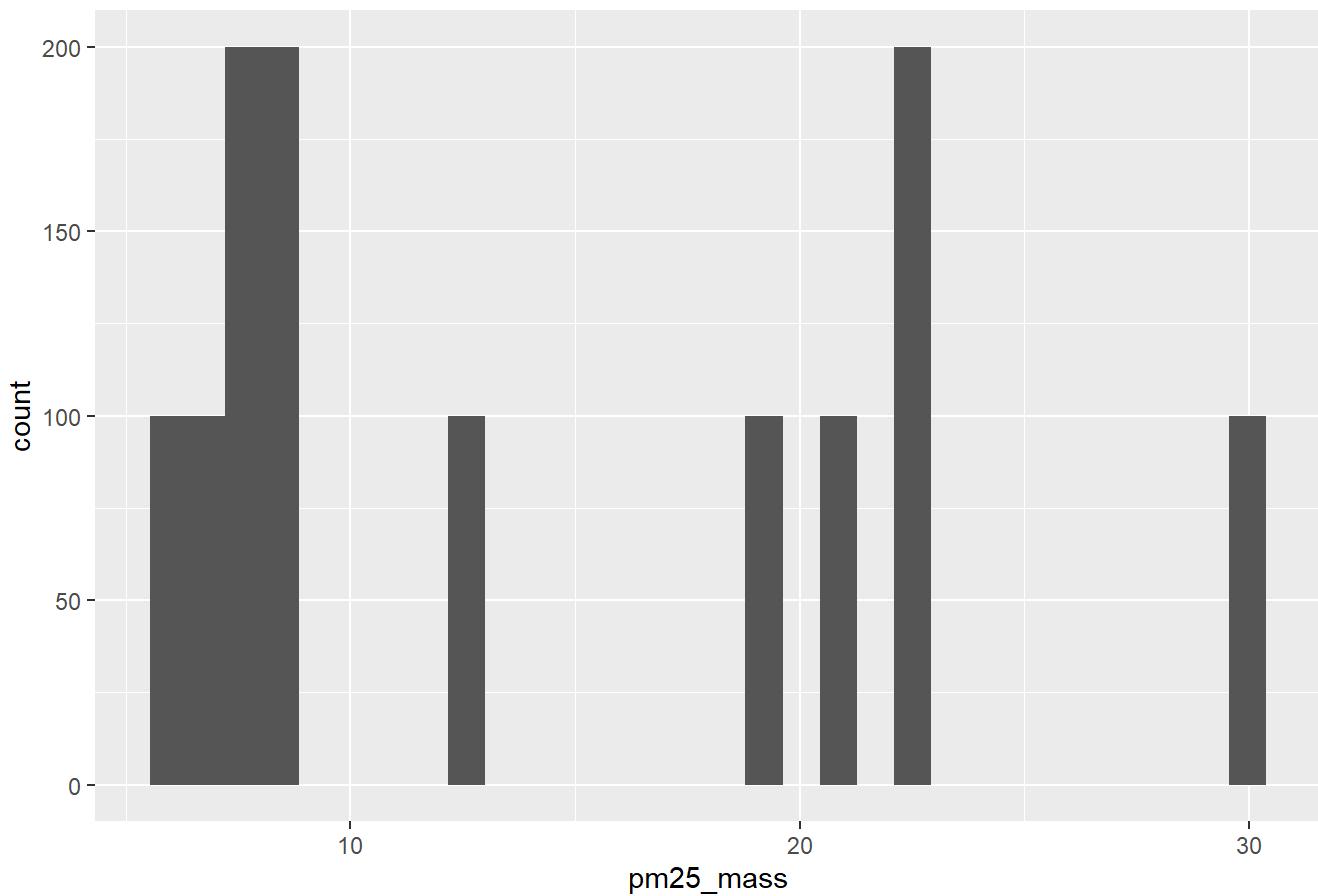
## Distribution of BMI



```
# PM2.5 mass (community-level, repeats by child after merge)
ggplot(chs, aes(x = pm25_mass)) + geom_histogram() + labs(title = "Distribution of PM2.5 (pm25_ma
```

`stat\_bin()` using `bins = 30`. Pick better value `binwidth`.

### Distribution of PM2.5 (pm25\_mass)



```
# Smoke and gas stove
chs %>% count(smoke, useNA = "ifany")
```

	smoke	useNA	n
1	0	ifany	970
2	1	ifany	190
3	NA	ifany	40

```
chs %>% count(gassstove, useNA = "ifany")
```

	gassstove	useNA	n
1	0	ifany	255
2	1	ifany	912
3	NA	ifany	33

From these histograms, we see that FEV follows a unimodal distribution centered around  $\sim 1900\text{-}2200$  ml. It is roughly centered around the mean without being clearly left-tailed or right-tailed. The BMI distribution is clearly right-skewed, with most values clustered around  $\sim 14\text{-}20$  followed by a long tail. Finally, in the distribution of PM2.5, it is not smooth, but instead contains one "spike" for each town, since PM2.5 is measured at the town/community level.

## Step 6: Check your expectations

```
# FEV by sex (male=1 usually higher)
chs %>%
  mutate(sex = factor(male, levels = c(0,1), labels = c("female","male"))) %>%
  group_by(sex) %>%
  summarise(mean_fev = mean(fev, na.rm=TRUE), sd_fev = sd(fev, na.rm=TRUE), n = sum(!is.na(fev)))
```

```
# A tibble: 2 × 4
  sex     mean_fev sd_fev     n
  <fct>    <dbl>   <dbl> <int>
1 female    1959.    327.   554
2 male      2104.    318.   551
```

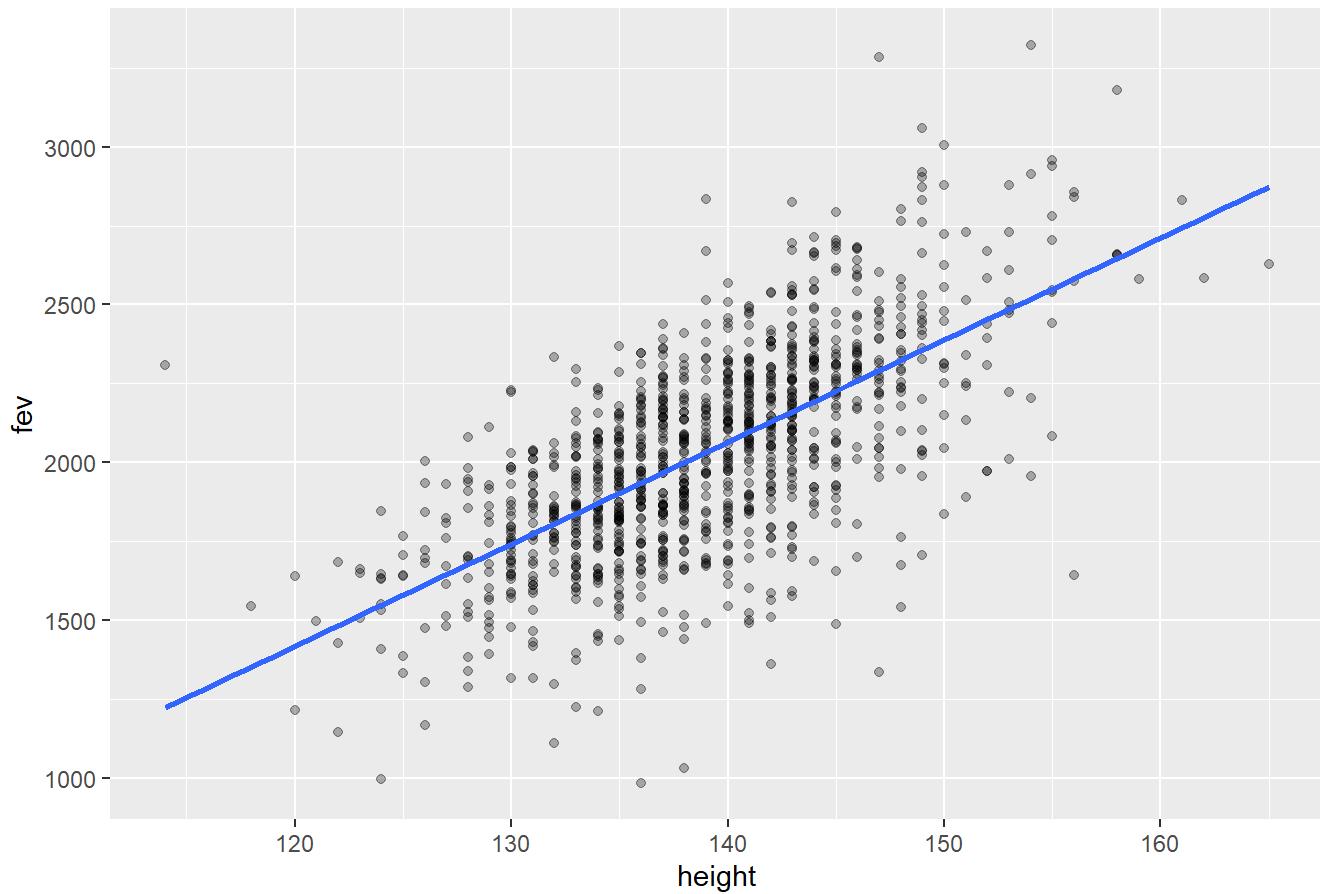
```
# FEV vs height (should increase)
ggplot(chs, aes(x = height, y = fev)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "FEV vs Height (sanity check)")
```

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

Warning: Removed 95 rows containing non-finite outside the scale range  
(`stat\_smooth()`).

Warning: Removed 95 rows containing missing values or values outside the scale range  
(`geom\_point()`).

## FEV vs Height (sanity check)



## Steps 7 and 8: Validate with an external source + Formulate a (simple) question

We ask the following question: **Do the patterns match basic physiology?**

If FEV increases with height and is higher in males, then the data is internally consistent, since taller individuals tend to have larger lung capacities.

**Source:** Almaasfeh S, Abukonna A, Omer S, Osman H. Evaluation of Forced Expiratory Volume in One Second and Forced Vital Capacity from Age and Height for Pulmonary Function Test. West Afr J Med. 2023 Oct 31;40(10):1029-1034. PMID: 37906250.

## Steps 9 and 10: Try the easy solution first + Challenge your solution

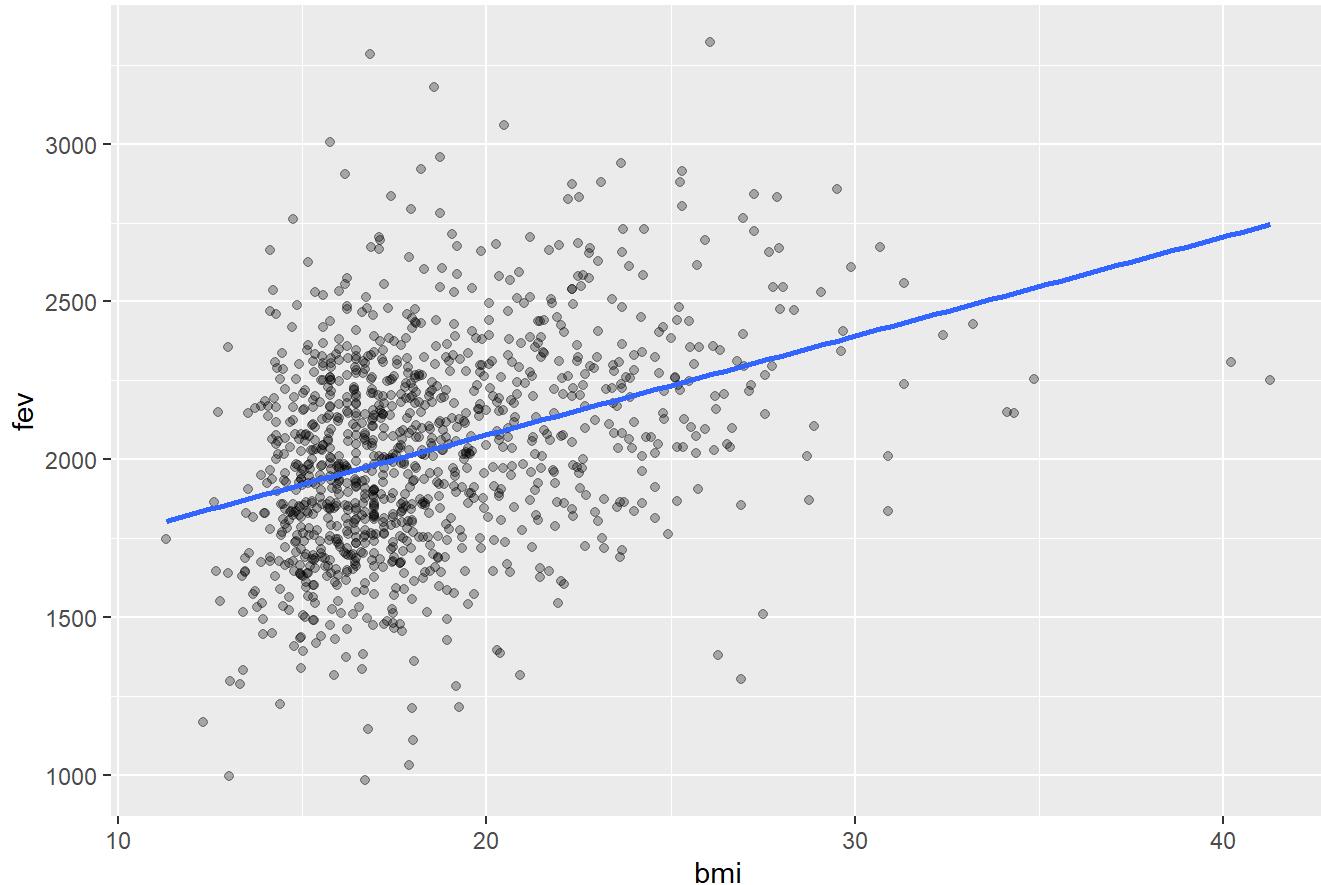
### Q1: Association between BMI and FEV

```
df1 <- chs %>% filter(!is.na(bmi), !is.na(fev))
```

```
ggplot(df1, aes(x = bmi, y = fev)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "FEV vs BMI")
```

`geom\_smooth()` using formula = 'y ~ x'

### FEV vs BMI



```
m1_easy <- lm(fev ~ bmi, data = df1)
summary(m1_easy)
```

Call:

`lm(formula = fev ~ bmi, data = df1)`

Residuals:

Min	1Q	Median	3Q	Max
-991.5	-207.9	-5.4	200.9	1304.8

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1450.760	46.497	31.20	<2e-16 ***
bmi	31.363	2.461	12.74	<2e-16 ***
---				

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 308.9 on 1103 degrees of freedom  
 Multiple R-squared: 0.1283, Adjusted R-squared: 0.1275  
 F-statistic: 162.4 on 1 and 1103 DF, p-value: < 2.2e-16

```
# Challenge (adjust for sex/height/age)
m1_adj <- lm(fev ~ bmi + male + height + agepft, data = df1)
summary(m1_adj)
```

Call:

```
lm(formula = fev ~ bmi + male + height + agepft, data = df1)
```

Residuals:

Min	1Q	Median	3Q	Max
-970.0	-142.6	11.1	154.1	977.2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2370.4352	199.1887	-11.900	< 2e-16 ***
bmi	12.8052	2.0173	6.348	3.19e-10 ***
male	119.1256	14.3995	8.273	3.75e-16 ***
height	29.4857	1.1672	25.261	< 2e-16 ***
agepft	0.7716	17.3223	0.045	0.964

---

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 237.3 on 1100 degrees of freedom  
 Multiple R-squared: 0.4867, Adjusted R-squared: 0.4848  
 F-statistic: 260.8 on 4 and 1100 DF, p-value: < 2.2e-16

In the easy (unadjusted) model, BMI is strongly and positively associated with lung function: each 1-unit increase in BMI corresponds to about 31.4 ml higher FEV ( $p < 2e-16$ ), and BMI alone explains a modest share of FEV variability ( $R^2 \approx 0.13$ ). In the challenged (adjusted) model controlling for sex, height, and age, the BMI association shrinks substantially to about 12.8 ml higher FEV per BMI unit, but remains highly statistically significant ( $p \approx 3.2e-10$ ). This drop in the BMI coefficient suggests that a large part of the crude BMI-FEV relationship reflects body size/physiology captured by height and sex (both strongly predictive:  $\sim +29.5$  ml per height unit and  $\sim +119$  ml for males), rather than BMI alone.

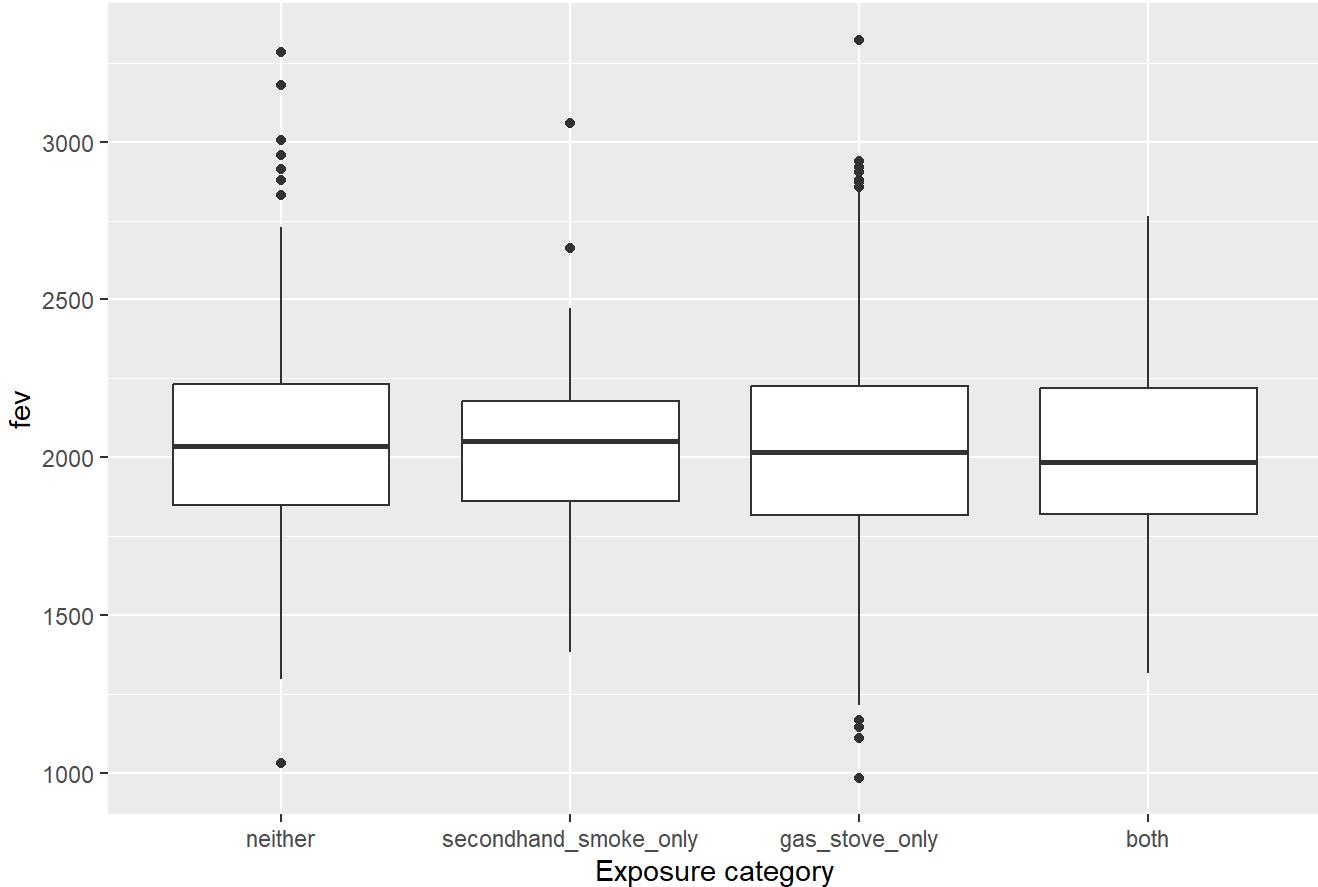
After adjustment, the model's explanatory power increases dramatically ( $R^2 \approx 0.49$ ), indicating that height and sex account for much more variation in FEV than BMI does; nonetheless, BMI still shows an independent positive association with FEV in this dataset, while age contributes little here ( $p \approx 0.96$ ).

## Q2: Association between smoke/gas exposure and FEV

```
df2 <- chs_imp %>%
  filter(!is.na(smoke_gas_exposure), !is.na(fev))
```

```
ggplot(df2, aes(x = smoke_gas_exposure, y = fev)) +
  geom_boxplot() +
  labs(title = "FEV by smoke/gas exposure", x = "Exposure category")
```

### FEV by smoke/gas exposure



```
df2 %>%
  group_by(smoke_gas_exposure) %>%
  summarise(mean_fev = mean(fev), sd_fev = sd(fev), n = n(), .groups = "drop")
```

```
# A tibble: 4 × 4
  smoke_gas_exposure    mean_fev   sd_fev     n
  <fct>                  <dbl>    <dbl>   <int>
1 neither                2055.    330.    214
2 secondhand_smoke_only  2056.    296.     36
3 gas_stove_only         2026.    318.    739
4 both                   2020.    299.    151
```

```
# Challenge (adjust for sex/height/age)
raw_tab <- df2 %>%
  group_by(smoke_gas_exposure) %>%
  summarise(
    mean_fev = mean(fev),
    sd_fev = sd(fev),
```

```

n = n(),
.groups = "drop"
)

m2_adj <- lm(fev ~ smoke_gas_exposure + male + height + agepft, data = df2)
summary(m2_adj)

```

Call:

```
lm(formula = fev ~ smoke_gas_exposure + male + height + agepft,
  data = df2)
```

Residuals:

Min	1Q	Median	3Q	Max
-910.42	-139.58	1.12	141.50	992.57

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2366.938	200.452	-11.808	<2e-16
smoke_gas_exposuresecondhand_smoke_only	-19.348	41.594	-0.465	0.642
smoke_gas_exposuregas_stove_only	-1.479	17.954	-0.082	0.934
smoke_gas_exposureboth	-27.043	24.569	-1.101	0.271
male	130.595	13.782	9.476	<2e-16
height	31.471	1.086	28.971	<2e-16
agepft	-3.537	17.439	-0.203	0.839
(Intercept)	***			
smoke_gas_exposuresecondhand_smoke_only				
smoke_gas_exposuregas_stove_only				
smoke_gas_exposureboth				
male	***			
height	***			
agepft				
---				
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Residual standard error: 230.8 on 1133 degrees of freedom

Multiple R-squared: 0.4723, Adjusted R-squared: 0.4695

F-statistic: 169 on 6 and 1133 DF, p-value: < 2.2e-16

```
library(emmeans)
```

Welcome to emmeans.

Caution: You lose important information if you filter this package's results.

See '? untidy'

```

adj_tab <- emmeans(m2_adj, ~ smoke_gas_exposure)
adj_tab

```

smoke_gas_exposure	emmmean	SE	df	lower.CL	upper.CL
neither	2038	15.8	1133	2007	2069
secondhand_smoke_only	2018	38.5	1133	1943	2094
gas_stove_only	2036	8.5	1133	2020	2053
both	2011	18.8	1133	1974	2048

Results are averaged over the levels of: male

Confidence level used: 0.95

```
adj_diff <- contrast(adj_tab, method = "trt.vs.ctrl", ref = "neither")
adj_diff
```

contrast	estimate	SE	df	t.ratio	p.value
secondhand_smoke_only - neither	-19.35	41.6	1133	-0.465	0.9098
gas_stove_only - neither	-1.48	18.0	1133	-0.082	0.9974
both - neither	-27.04	24.6	1133	-1.101	0.5469

Results are averaged over the levels of: male

P value adjustment: dunnettx method for 3 tests

```
raw_only <- raw_tab %>%
  select(smoke_gas_exposure, raw_mean = mean_fev, raw_sd = sd_fev, n)

adj_only <- as.data.frame(adj_tab) %>%
  select(smoke_gas_exposure, adj_mean = emmean, adj_se = SE, df)

comparison <- left_join(raw_only, adj_only, by = "smoke_gas_exposure")
comparison
```

# A tibble: 4 × 7	smoke_gas_exposure	raw_mean	raw_sd	n	adj_mean	adj_se	df
1	neither	2055.	330.	214	2038.	15.8	1133
2	secondhand_smoke_only	2056.	296.	36	2018.	38.5	1133
3	gas_stove_only	2026.	318.	739	2036.	8.50	1133
4	both	2020.	299.	151	2011.	18.8	1133

In unadjusted comparisons, mean FEV was ~30–35 ml lower in the gas-stove-only and both-exposure groups compared with neither exposure

After adjusting for sex, height, and age, the difference for gas-stove-only essentially vanished (~29 ml unadjusted vs ~1.5 ml adjusted), and the difference for both exposures was smaller (about ~27 ml). None of the adjusted differences were statistically significant.

Height and sex were strongly associated with FEV, suggesting the crude differences largely reflected differences in body size/sex composition across exposure groups rather than a clear exposure effect.

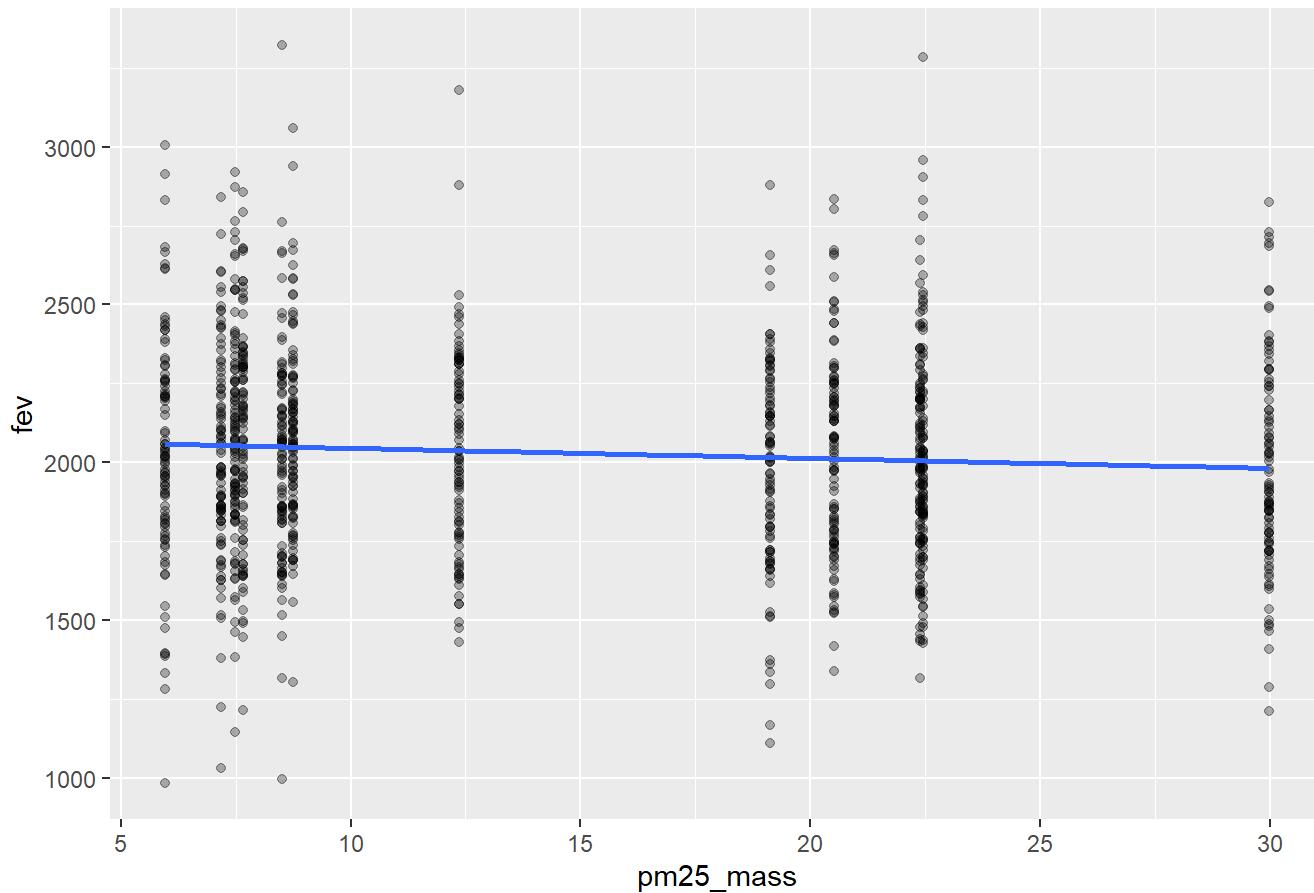
### Q3: Association between PM2.5 exposure and FEV

```
df3 <- chs %>% filter(!is.na(pm25_mass), !is.na(fev))

ggplot(df3, aes(x = pm25_mass, y = fev)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "FEV vs PM2.5 (pm25_mass)")

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

FEV vs PM2.5 (pm25\_mass)



```
m3_easy <- lm(fev ~ pm25_mass, data = df3)
summary(m3_easy)
```

Call:

```
lm(formula = fev ~ pm25_mass, data = df3)
```

Residuals:

Min	1Q	Median	3Q	Max
-1073.46	-224.99	-7.52	213.52	1277.68

Coefficients:

Estimate	Std. Error	t value	Pr(> t )
----------	------------	---------	----------

```
(Intercept) 2077.317      21.000  98.922   <2e-16 ***
pm25_mass    -3.189       1.282   -2.488    0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 329.9 on 1103 degrees of freedom  
 Multiple R-squared: 0.005583, Adjusted R-squared: 0.004681  
 F-statistic: 6.192 on 1 and 1103 DF, p-value: 0.01298

```
m3_adj <- lm(fev ~ pm25_mass + male + height + agepft, data = df3)
summary(m3_adj)
```

Call:

```
lm(formula = fev ~ pm25_mass + male + height + agepft, data = df3)
```

Residuals:

Min	1Q	Median	3Q	Max
-994.73	-148.80	10.95	155.82	1019.66

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2317.8960	207.5833	-11.166	<2e-16 ***
pm25_mass	-1.6473	0.9521	-1.730	0.0839 .
male	124.6853	14.6269	8.524	<2e-16 ***
height	31.9878	1.1149	28.691	<2e-16 ***
agepft	-13.5594	17.7876	-0.762	0.4460

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 241.3 on 1100 degrees of freedom  
 Multiple R-squared: 0.4694, Adjusted R-squared: 0.4674  
 F-statistic: 243.2 on 4 and 1100 DF, p-value: < 2.2e-16

In the easy (unadjusted) model, PM2.5 mass is negatively associated with lung function: each 1-unit increase in pm25\_mass is associated with about 3.19 ml lower FEV ( $p = 0.013$ ). However, this model explains essentially none of the variability in FEV ( $R^2 \approx 0.006$ ), meaning PM2.5 alone is a very weak predictor at the individual level.

In the challenged (adjusted) model controlling for sex, height, and age, the PM2.5 association becomes smaller in magnitude (about -1.65 ml per unit) and is only marginal ( $p = 0.084$ ), suggesting the crude association was at least partly confounded by differences in body size/sex/age across towns. Meanwhile, the adjustment variables behave as expected and dominate prediction: height (~+32 ml per unit) and male sex (~+125 ml) are strongly associated with higher FEV, and the model's  $R^2$  jumps to ~0.47 largely due to these physiological predictors rather than PM2.

### 3. Data visualizations

## (i) Facet plot showing scatterplots with regression lines of BMI vs FEV by "townname"

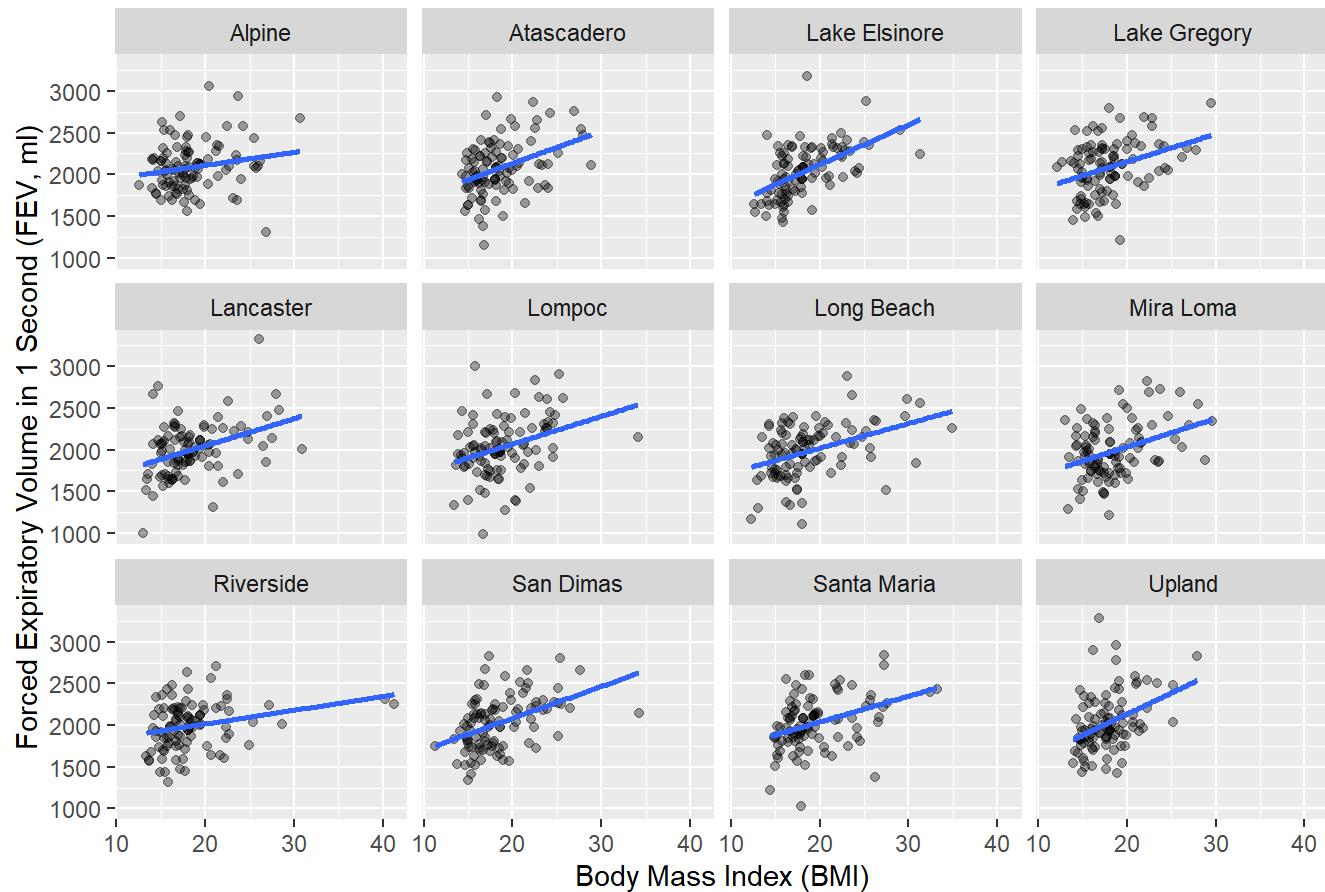
```
library(dplyr)
library(ggplot2)

df_bmi_fev <- chs_imp %>%
  filter(!is.na(bmi), !is.na(fev), !is.na(townname))

ggplot(df_bmi_fev, aes(x = bmi, y = fev)) +
  geom_point(alpha = 0.35) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ townname) +
  labs(
    title = "BMI vs FEV by Town",
    x = "Body Mass Index (BMI)",
    y = "Forced Expiratory Volume in 1 Second (FEV, ml)"
  )
```

`geom\_smooth()` using formula = 'y ~ x'

BMI vs FEV by Town



Across all 12 towns, the regression line slopes are positive, meaning higher BMI tends to be associated with higher FEV within each town. The strength of the relationship varies: some towns (e.g., Lake Elsinore, San

Dimas, Upland) show a steeper upward trend, while others (e.g., Riverside, Alpine) look flatter, suggesting a weaker within-town BMI–FEV association there. In every panel, there's also substantial scatter around the line, so BMI alone doesn't explain most of the person-to-person variation in FEV within a town. Finally, the cloud of points shifts up/down across towns, implying that towns differ in average FEV levels (possibly due to differences in age/height/sex composition or other town-level factors), even though the within-town BMI–FEV trend is generally consistent.

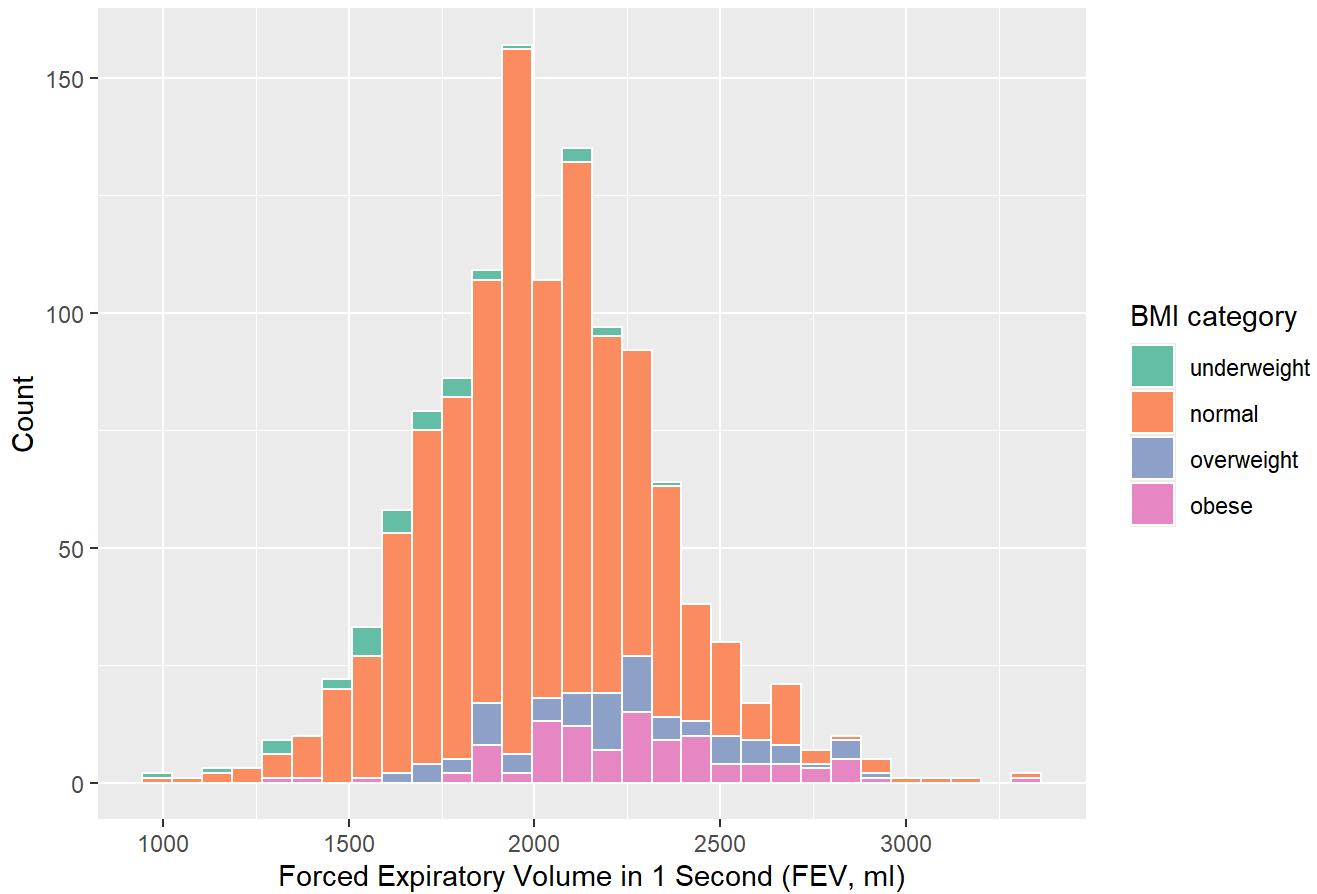
## (ii) Stacked histograms of FEV by BMI category and FEV by smoke/gas exposure

```
library(dplyr)
library(ggplot2)

df_plot <- chs_imp %>%
  filter(!is.na(fev))

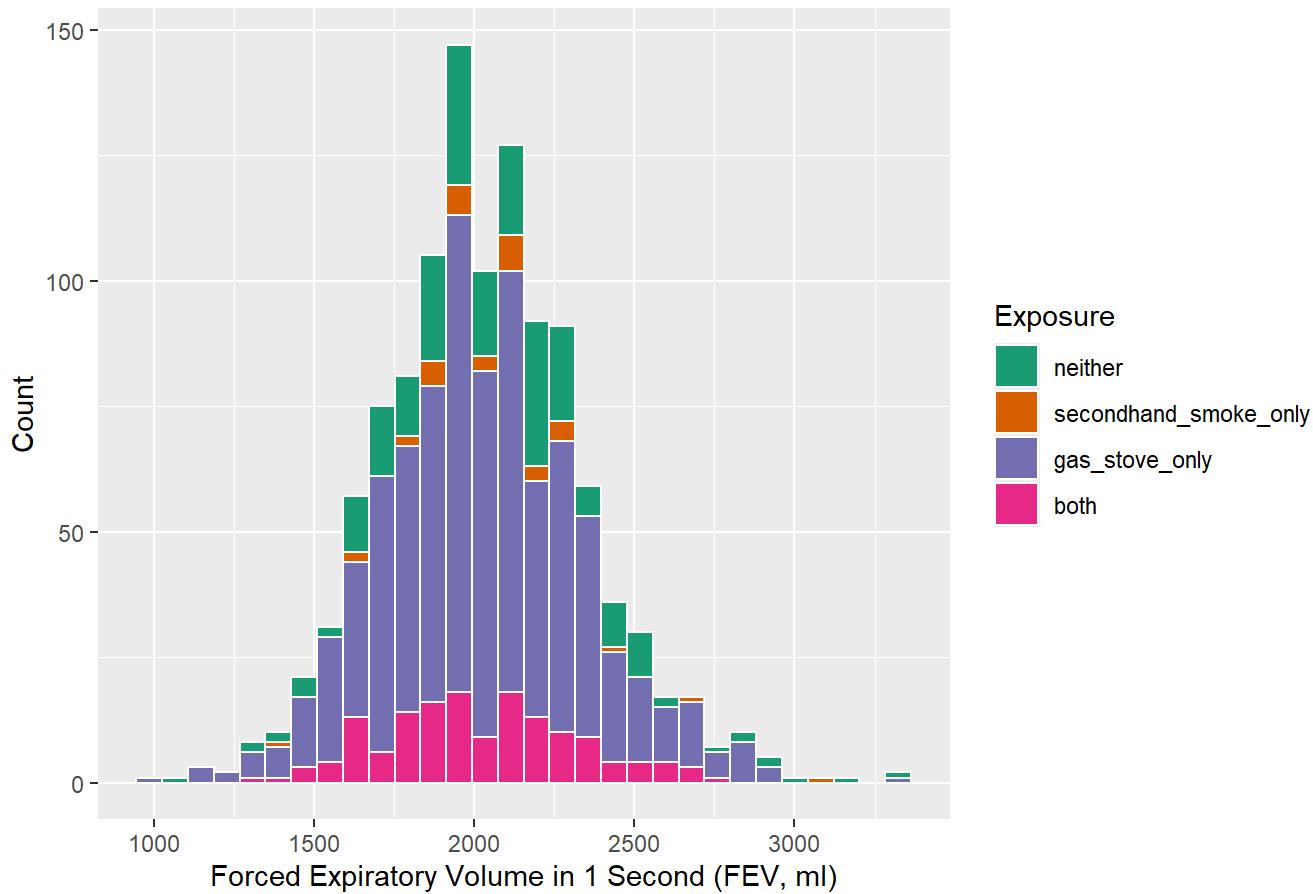
# 1) Stacked histogram: FEV by BMI category (obesity_level)
ggplot(df_plot %>% filter(!is.na(obesity_level)),
       aes(x = fev, fill = obesity_level)) +
  geom_histogram(position = "stack", bins = 30, color = "white") +
  scale_fill_brewer(palette = "Set2", name = "BMI category") +
  labs(
    title = "Distribution of FEV by BMI Category",
    x = "Forced Expiratory Volume in 1 Second (FEV, ml)",
    y = "Count"
  )
```

## Distribution of FEV by BMI Category



```
# 2) Stacked histogram: FEV by smoke/gas exposure
ggplot(df_plot %>% filter(!is.na(smoke_gas_exposure)),
       aes(x = fev, fill = smoke_gas_exposure)) +
  geom_histogram(position = "stack", bins = 30, color = "white") +
  scale_fill_brewer(palette = "Dark2", name = "Exposure") +
  labs(
    title = "Distribution of FEV by Smoke/Gas Exposure",
    x = "Forced Expiratory Volume in 1 Second (FEV, ml)",
    y = "Count"
  )
```

### Distribution of FEV by Smoke/Gas Exposure



FEV is roughly unimodal, BMI is right-skewed with a few high outliers, and PM2.5 appears in discrete spikes because it's measured at the town level and repeats across children within towns. BMI is positively associated with FEV, but the effect shrinks after adjusting for sex/height/age, indicating much of the raw association reflects body size. Smoke/gas exposure groups have very similar FEV distributions and show no clear adjusted differences from the "neither" group. PM2.5 has a small negative unadjusted association with FEV that weakens after adjustment, so any effect looks modest relative to individual variability.

### (iii) Barchart of BMI by smoke/gas exposure

```
library(dplyr)
library(ggplot2)

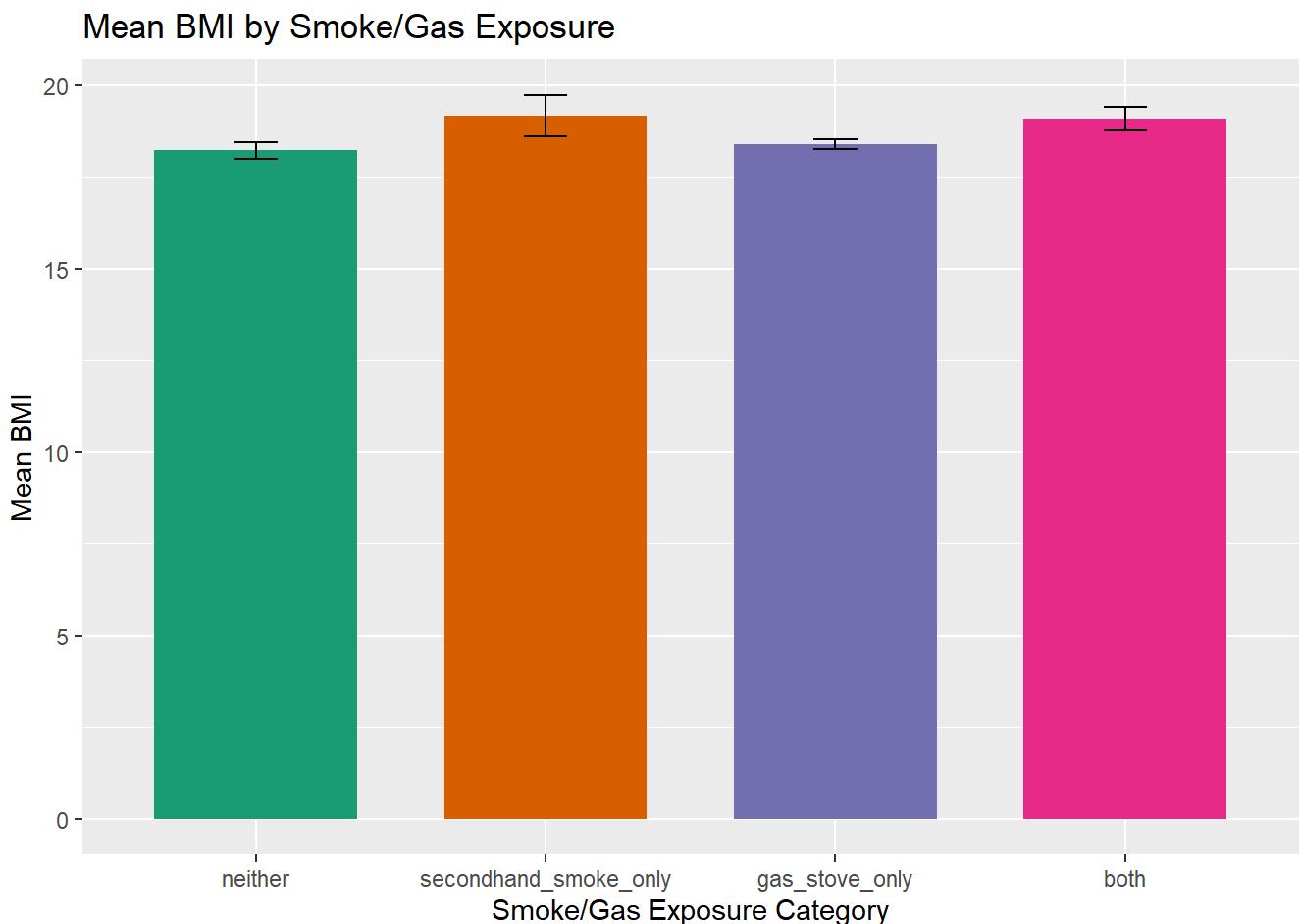
df_bmi_exposure <- chs_imp %>%
  filter(!is.na(bmi), !is.na(smoke_gas_exposure))

# Bar chart of mean BMI by exposure (with SE error bars)
bmi_sum <- df_bmi_exposure %>%
  group_by(smoke_gas_exposure) %>%
  summarise(
    mean_bmi = mean(bmi),
    sd_bmi = sd(bmi),
    n = n(),
```

```
    se_bmi = sd_bmi / sqrt(n),
    .groups = "drop"
)
bmi_sum
```

```
# A tibble: 4 × 5
  smoke_gas_exposure  mean_bmi   sd_bmi     n  se_bmi
  <fct>              <dbl>     <dbl>   <int>  <dbl>
1 neither             18.2      3.36    214  0.230
2 secondhand_smoke_only 19.2      3.38     36  0.564
3 gas_stove_only      18.4      3.65    739  0.134
4 both                19.1      3.97    151  0.323
```

```
ggplot(bmi_sum, aes(x = smoke_gas_exposure, y = mean_bmi, fill = smoke_gas_exposure)) +
  geom_col(width = 0.7) +
  geom_errorbar(aes(ymin = mean_bmi - se_bmi, ymax = mean_bmi + se_bmi), width = 0.15) +
  scale_fill_brewer(palette = "Dark2", guide = "none") +
  labs(
    title = "Mean BMI by Smoke/Gas Exposure",
    x = "Smoke/Gas Exposure Category",
    y = "Mean BMI"
)
```



Mean BMI is fairly similar across the four smoke/gas exposure groups, clustering around roughly 18–19.5. The secondhand\_smoke\_only and both groups have slightly higher mean BMI than neither, while gas\_stove\_only is close to neither. The error bars overlap across all groups, suggesting these differences are small and may not be practically meaningful (especially given the likely smaller sample size for secondhand\_smoke\_only).

#### (iv) Statistical summary graphs of FEV by BMI and FEV by smoke/gas exposure category

```
library(dplyr)
library(ggplot2)

df <- chs_imp %>%
  filter(!is.na(fev))

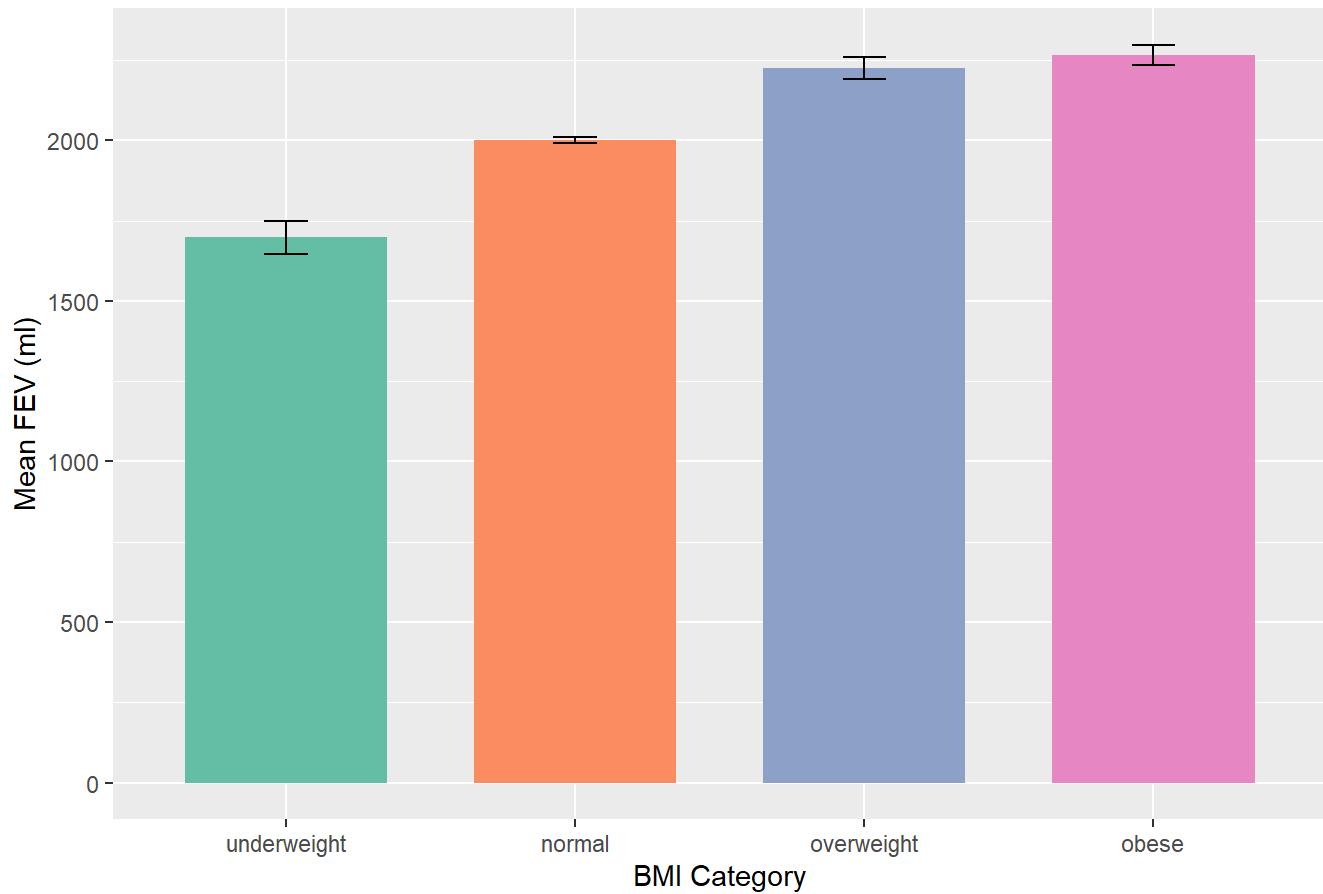
# 1) Statistical summary: FEV by BMI category (mean + SE)
sum_bmi <- df %>%
  filter(!is.na(obesity_level)) %>%
  group_by(obesity_level) %>%
  summarise(
    mean_fev = mean(fev),
    sd_fev = sd(fev),
    n = n(),
    se_fev = sd_fev / sqrt(n),
    .groups = "drop"
  )

sum_bmi

# A tibble: 4 × 5
  obesity_level mean_fev sd_fev     n se_fev
  <fct>        <dbl>   <dbl> <int>  <dbl>
1 underweight    1698.   303.    35   51.3
2 normal         2000.   295.   975   9.45
3 overweight     2224.   317.    87   34.0
4 obese          2266.   325.   103   32.1

ggplot(sum_bmi, aes(x = obesity_level, y = mean_fev, fill = obesity_level)) +
  geom_col(width = 0.7) +
  geom_errorbar(aes(ymin = mean_fev - se_fev, ymax = mean_fev + se_fev), width = 0.15) +
  scale_fill_brewer(palette = "Set2", guide = "none") +
  labs(
    title = "Mean FEV by BMI Category",
    x = "BMI Category",
    y = "Mean FEV (ml)"
  )
```

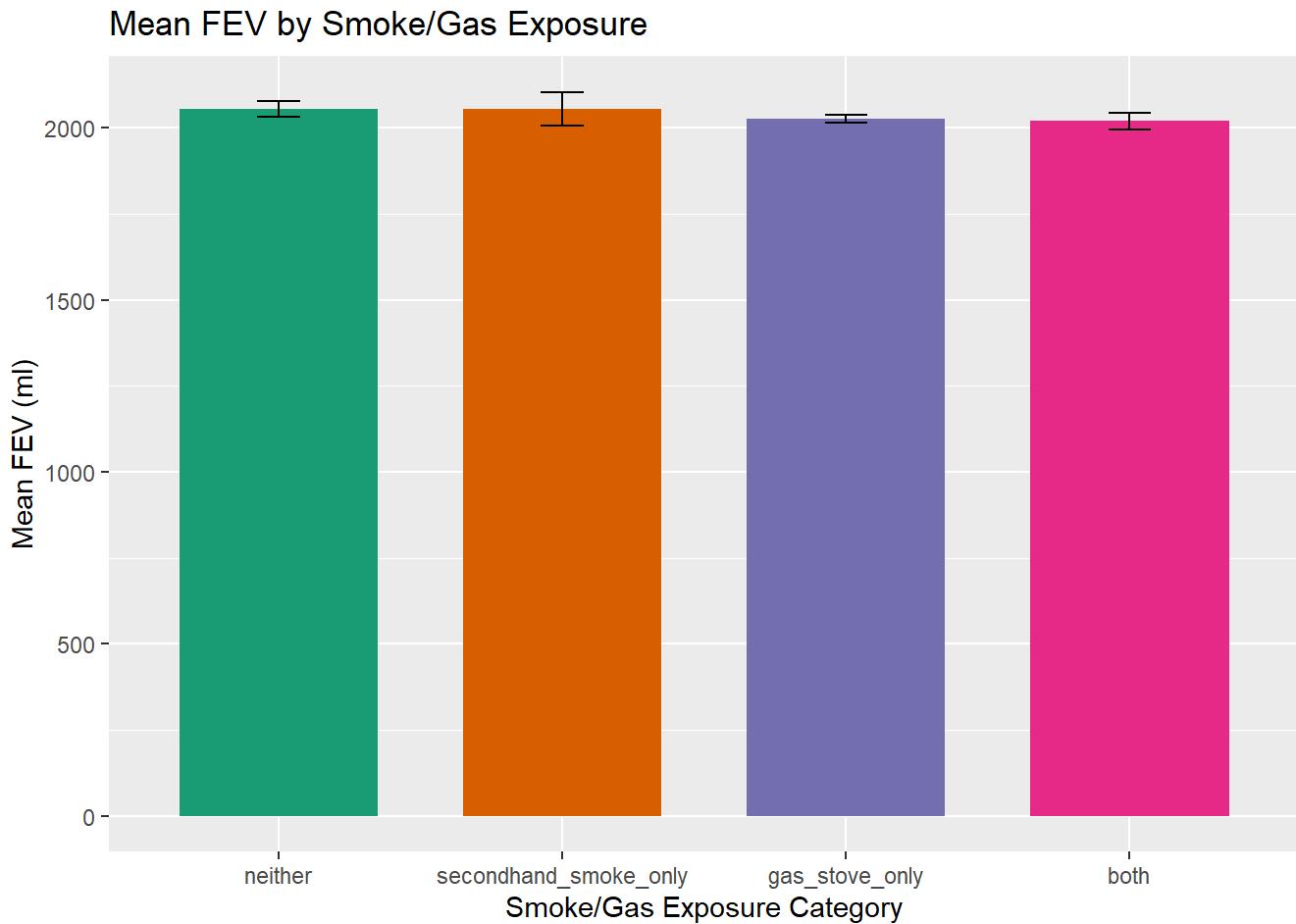
## Mean FEV by BMI Category



```
# 2) Statistical summary: FEV by smoke/gas exposure (mean + SE)
sum_exposure <- df %>%
  filter(!is.na(smoke_gas_exposure)) %>%
  group_by(smoke_gas_exposure) %>%
  summarise(
    mean_fev = mean(fev),
    sd_fev = sd(fev),
    n = n(),
    se_fev = sd_fev / sqrt(n),
    .groups = "drop"
  )
sum_exposure
```

```
# A tibble: 4 × 5
  smoke_gas_exposure  mean_fev  sd_fev     n  se_fev
  <fct>              <dbl>   <dbl> <int>  <dbl>
1 neither            2055.    330.   214   22.6
2 secondhand_smoke_only 2056.    296.    36   49.3
3 gas_stove_only      2026.    318.   739   11.7
4 both               2020.    299.   151   24.3
```

```
ggplot(sum_exposure, aes(x = smoke_gas_exposure, y = mean_fev, fill = smoke_gas_exposure)) +
  geom_col(width = 0.7) +
  geom_errorbar(aes(ymin = mean_fev - se_fev, ymax = mean_fev + se_fev), width = 0.15) +
  scale_fill_brewer(palette = "Dark2", guide = "none") +
  labs(
    title = "Mean FEV by Smoke/Gas Exposure",
    x = "Smoke/Gas Exposure Category",
    y = "Mean FEV (ml)"
  )
```



Mean FEV increases steadily across BMI categories, from lowest in the underweight group to highest in the overweight/obese groups, suggesting a clear positive association between body size and lung volume. The error bars are small relative to the differences between BMI groups, so this trend looks consistent. In contrast, mean FEV is very similar across the four smoke/gas exposure categories, with only small differences and largely overlapping error bars. Overall, BMI category shows a much stronger relationship with FEV than smoke/gas exposure does in these summaries.

## (v) A leaflet map showing the concentrations of PM2.5 mass in each of the CHS communities

```
library(dplyr)
library(leaflet)

chsRegional <- read.csv("chs_Regional.csv")
head(chsRegional)
```

	townname	pm25_mass	pm25_so4	pm25_no3	pm25_nh4	pm25_oc	pm25_ec	pm25_om				
1	Alpine	8.74	1.73	1.59	0.88	2.54	0.48	3.04				
2	Lake Elsinore	12.35	1.90	2.98	1.36	3.64	0.62	4.36				
3	Lake Gregory	7.66	1.07	2.07	0.91	2.46	0.40	2.96				
4	Lancaster	8.50	0.91	1.87	0.78	4.43	0.55	5.32				
5	Lompoc	5.96	1.08	0.73	0.41	1.45	0.13	1.74				
6	Long Beach	19.12	3.23	6.22	2.57	5.21	1.36	6.25				
		pm10_oc	pm10_ec	pm10_tc	formic	acetic	hcl	hno3	o3_max	o3106	o3_24	no2
1	3.25	0.49	3.75	1.03	2.49	0.41	1.98	65.82	55.05	41.23	12.18	
2	4.66	0.63	5.29	1.18	3.56	0.46	2.63	66.70	54.42	32.23	17.03	
3	3.16	0.41	3.57	0.66	2.36	0.28	2.28	84.44	67.01	57.76	7.62	
4	5.68	0.56	8.61	0.88	2.88	0.22	1.80	54.81	43.88	32.86	15.77	
5	1.86	0.14	1.99	0.34	0.75	0.33	0.43	43.85	37.74	28.37	4.60	
6	6.68	1.39	8.07	1.57	2.94	0.73	2.67	39.44	28.22	18.22	33.11	
		pm10_no_24hr	pm2_5_fr	iacid	oacid	total_acids		lon	lat			
1	24.73	2.48	10.28	2.39	3.52		5.50	-116.7664	32.83505			
2	34.25	7.07	14.53	3.09	4.74		7.37	-117.3273	33.66808			
3	20.05	NA	9.01	2.56	3.02		5.30	-117.2752	34.24290			
4	25.04	12.68	NA	2.02	3.76		5.56	-118.1542	34.68678			
5	18.40	2.05	NA	0.76	1.09		1.52	-120.4579	34.63915			
6	38.41	36.76	22.23	3.40	4.51		7.18	-118.1937	33.77005			

```
names(chsRegional)
```

[1]	"townname"	"pm25_mass"	"pm25_so4"	"pm25_no3"	"pm25_nh4"
[6]	"pm25_oc"	"pm25_ec"	"pm25_om"	"pm10_oc"	"pm10_ec"
[11]	"pm10_tc"	"formic"	"acetic"	"hcl"	"hno3"
[16]	"o3_max"	"o3106"	"o3_24"	"no2"	"pm10"
[21]	"no_24hr"	"pm2_5_fr"	"iacid"	"oacid"	"total_acids"
[26]	"lon"	"lat"			

```
# PM2.5 mass column
pm_var <- "pm25_mass"

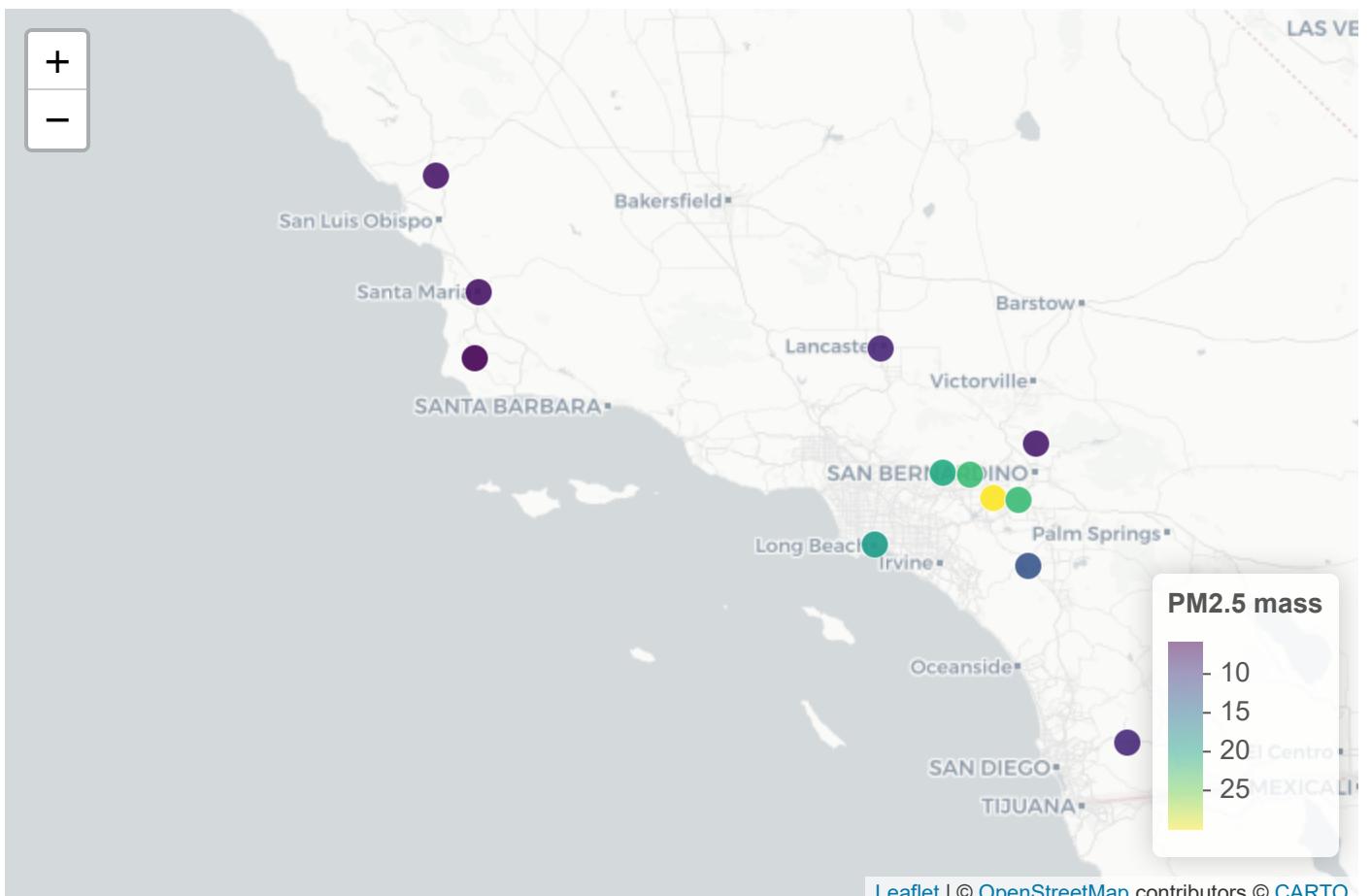
# Quick check
chsRegional %>%
  select(townname, all_of("lat"), all_of("lon"), all_of(pm_var)) %>%
  head()
```

	townname	lat	lon	pm25_mass
1	Alpine	32.83505	-116.7664	8.74
2	Lake Elsinore	33.66808	-117.3273	12.35
3	Lake Gregory	34.24290	-117.2752	7.66

4	Lancaster	34.68678	-118.1542	8.50
5	Lompoc	34.63915	-120.4579	5.96
6	Long Beach	33.77005	-118.1937	19.12

```
pal <- colorNumeric(palette = "viridis", domain = chsRegional[[pm_var]])

leaflet(chsRegional) %>%
  addProviderTiles(providers$CartoDB.Positron) %>%
  addCircleMarkers(
    lng = ~lon,
    lat = ~lat,
    fillColor = ~pal(pm25_mass),
    fillOpacity = 0.9,
    radius = 7,
    stroke = TRUE,
    weight = 1,
    color = "white",
    popup = ~paste0("<b>", townname, "</b><br>PM2.5 mass: ", round(pm25_mass, 2))
  ) %>%
  addLegend("bottomright", pal = pal, values = ~pm25_mass, title = "PM2.5 mass")
```



Leaflet | © OpenStreetMap contributors © CARTO

The leaflet map shows PM2.5 mass levels for each CHS community, with dot color indicating concentration according to the legend: purple dots represent lower PM2.5 (around 10), colors transition through blue/green for mid-range values (~15–20), and yellow dots indicate the highest PM2.5 (around 25). Most of the highest-PM2.5 communities (yellow/green) cluster in the inland/greater Los Angeles area, while several

coastal or more western communities appear in the lower range (purple). Overall, the map highlights substantial geographic variation in PM2.5 exposure across communities, supporting town-level PM2.5 as a meaningful exposure variable in the FEV analyses.

## (vi) Choose a visualization to examine whether PM2.5 mass is associated with FEV

```
library(dplyr)
library(ggplot2)

df3 <- chs_imp %>%
  filter(!is.na(fev), !is.na(pm25_mass), !is.na(townname))

# (Recommended) Town-level visualization:
# Plot the mean FEV in each town vs the town's PM2.5, with a regression line.
town_summary <- df3 %>%
  group_by(townname, pm25_mass) %>%
  summarise(
    mean_fev = mean(fev),
    n = n(),
    .groups = "drop"
  )

town_summary
```

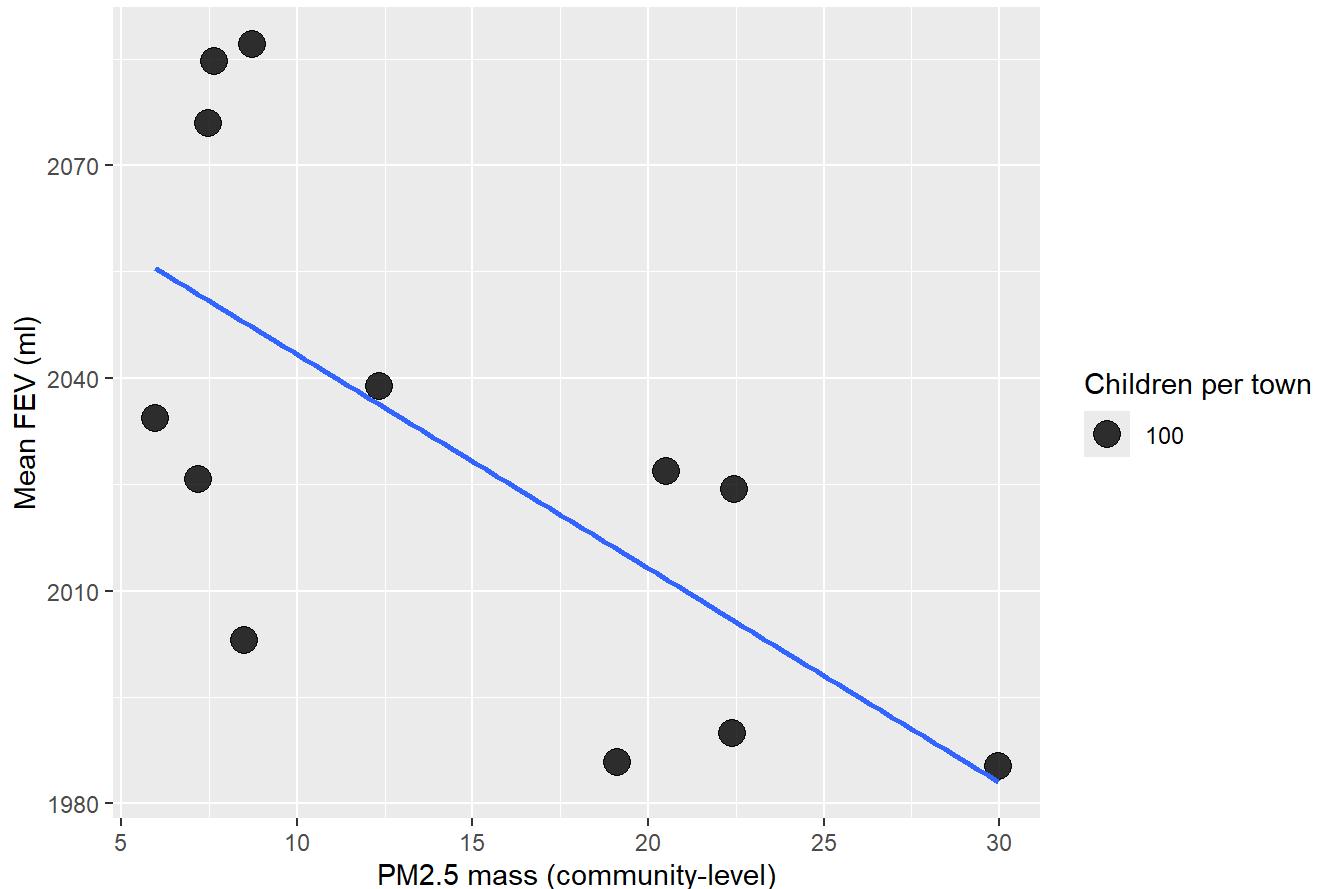
```
# A tibble: 12 × 4
  townname      pm25_mass mean_fev     n
  <chr>          <dbl>     <dbl> <int>
1 Alpine         8.74     2087.   100
2 Atascadero    7.48     2076.   100
3 Lake Elsinore 12.4     2039.   100
4 Lake Gregory   7.66     2085.   100
5 Lancaster      8.5      2003.   100
6 Lompoc         5.96     2034.   100
7 Long Beach     19.1     1986.   100
8 Mira Loma      30.0     1985.   100
9 Riverside       22.4     1990.   100
10 San Dimas     20.5     2027.   100
11 Santa Maria    7.19     2026.   100
12 Upland         22.5     2024.   100
```

```
ggplot(town_summary, aes(x = pm25_mass, y = mean_fev)) +
  geom_point(aes(size = n), alpha = 0.8) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Mean FEV vs PM2.5 Mass by Community",
    x = "PM2.5 mass (community-level)",
    y = "Mean FEV (ml)",
```

```
size = "Children per town"
)
```

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

### Mean FEV vs PM2.5 Mass by Community



This community-level plot shows a negative association between PM2.5 mass and mean FEV: towns with higher PM2.5 tend to have lower average FEV, as indicated by the downward regression line. However, there is noticeable scatter around the line, so PM2.5 alone does not perfectly explain differences in mean FEV across towns. Because each point represents a town average (with size reflecting the number of children), the pattern reflects between-community differences rather than individual-level variation. Overall, the figure is consistent with a modest adverse relationship between PM2.5 exposure and lung function in this dataset.