Final Project

Karina Curiel

2022-09-19

Introduction

This data come from a study on chronic respiratory disease and the effects of various types of smog on lung function of children and adults in the Los Angeles area.

```
lung <- read.delim("/Users/karinacuriel/Desktop/math 130/data/Lung_081217.txt", header=TRUE,sep="\t")
```

I want to look into the lung function of the oldest child in one area versus the lung function of the oldest child in another area. How does lung function vary based on location and oldest children?

Univariate Exploration:

```
summary(lung$OCAGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    7.00   10.00   13.00   12.63   16.00   17.00
```

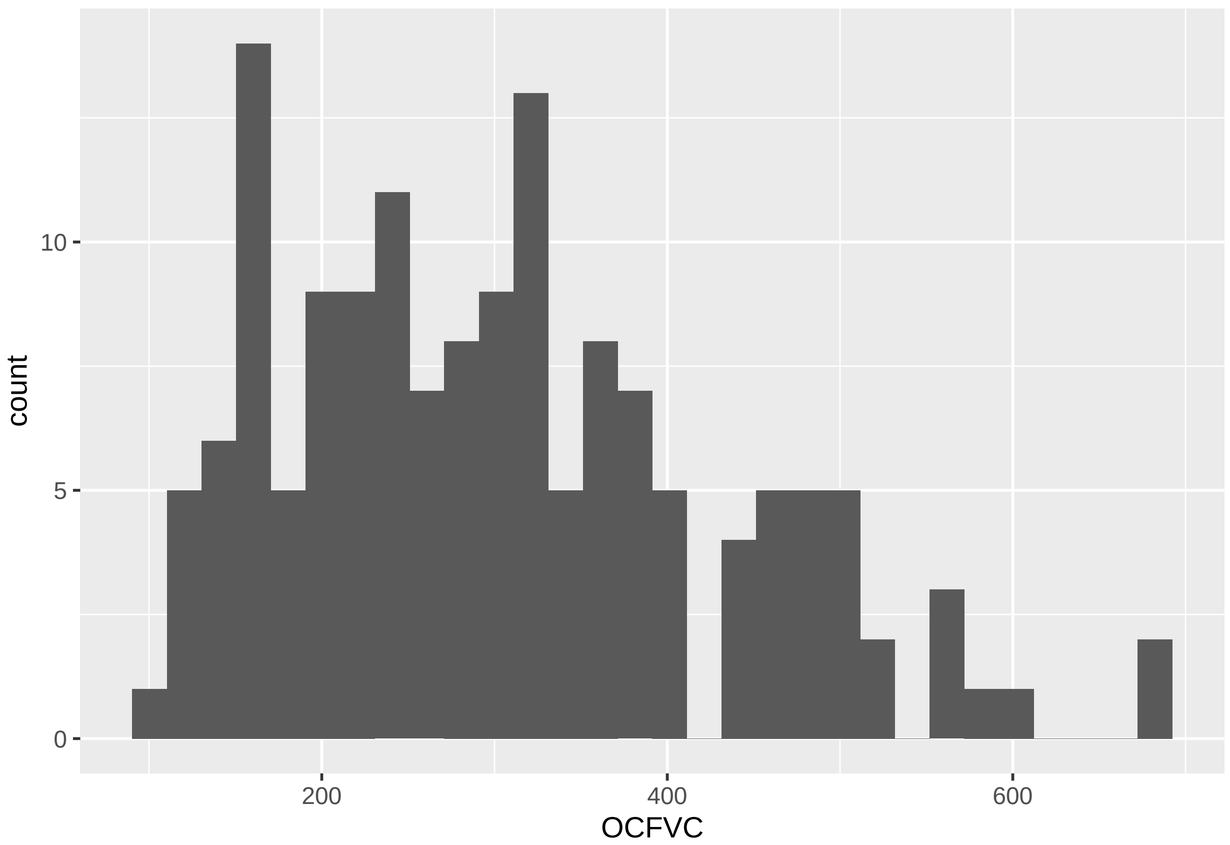
```
summary(lung$OCFVC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   107.0   201.0   291.0   304.3   376.0   689.0
```

This summary data shows the range of the age of the oldest child as well as the range for the forced expiratory volume which measures how much air a person can exhale during a forced breath.

```
library(ggplot2)
ggplot(lung, aes(x=OCFVC)) + geom_histogram()
```

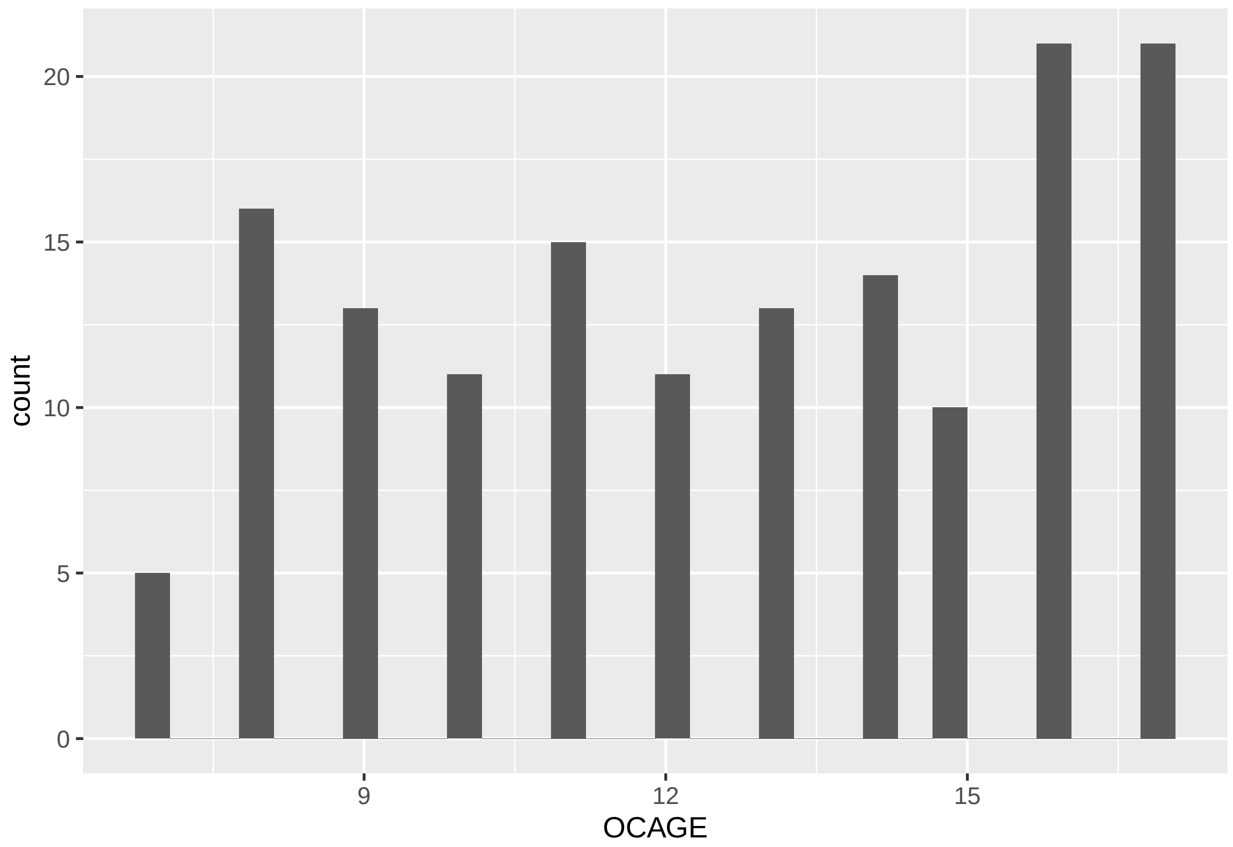
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The data is skewed to the left and shows the amount of air forced out during an exhale. In general the breath volume does not gather much past 400.

```
library(ggplot2)
ggplot(lung, aes(x=OCAGE)) + geom_histogram()
```

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



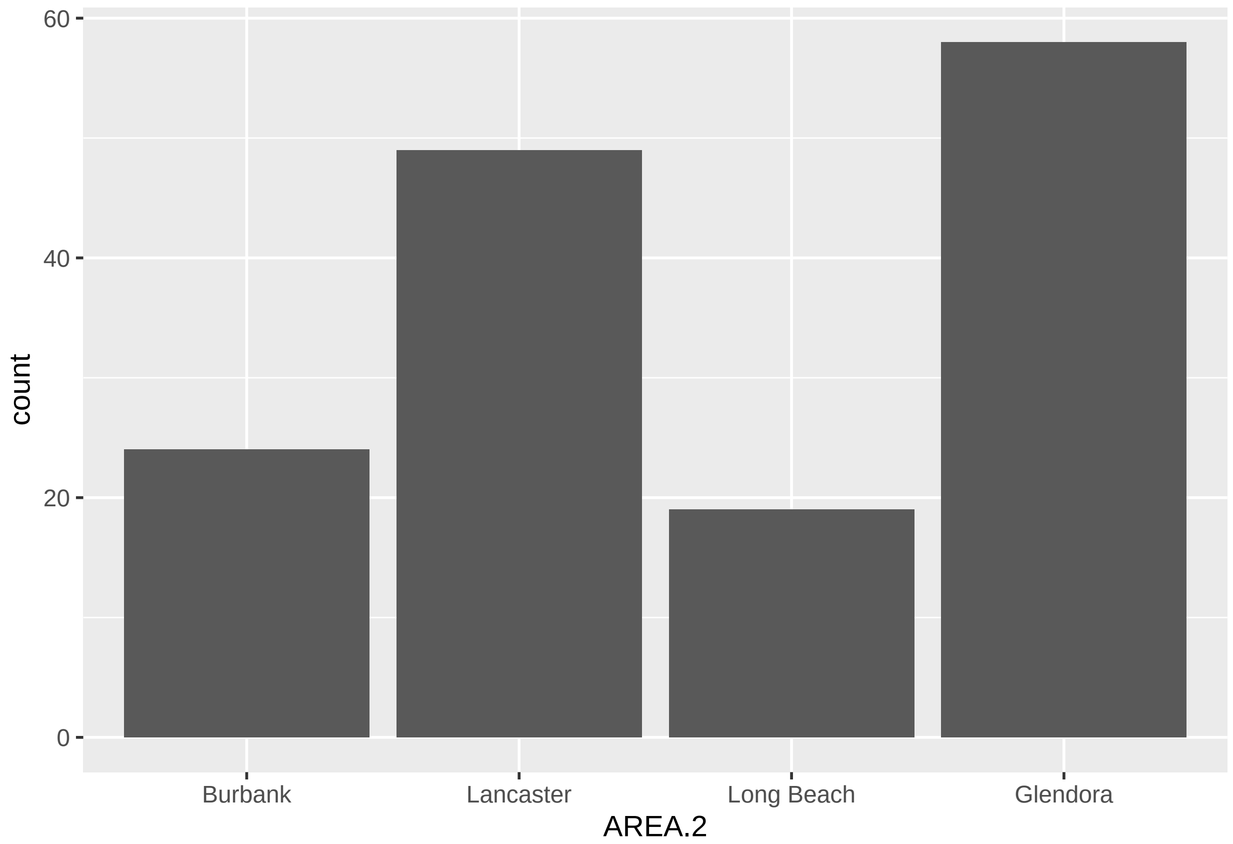
```
library(dplyr)
```

```
##
## 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
```

```
lung$AREA.2 <- factor(lung$AREA, labels=c( "Burbank", "Lancaster","Long Beach", "Glendora"))
table(lung$AREA.2, lung$AREA, useNA="always")
```

```
##
##                 1  2  3  4 <NA>
##   Burbank      24  0  0  0    0
##   Lancaster     0 49  0  0    0
##   Long Beach    0  0 19  0    0
##   Glendora      0  0  0 58    0
##   <NA>          0  0  0  0    0
```

```
ggplot(lung, aes(x=AREA.2)) + geom_bar()
```



Bivariate Exploration

Grouped barcahrt

```
library(ggplot2)
```

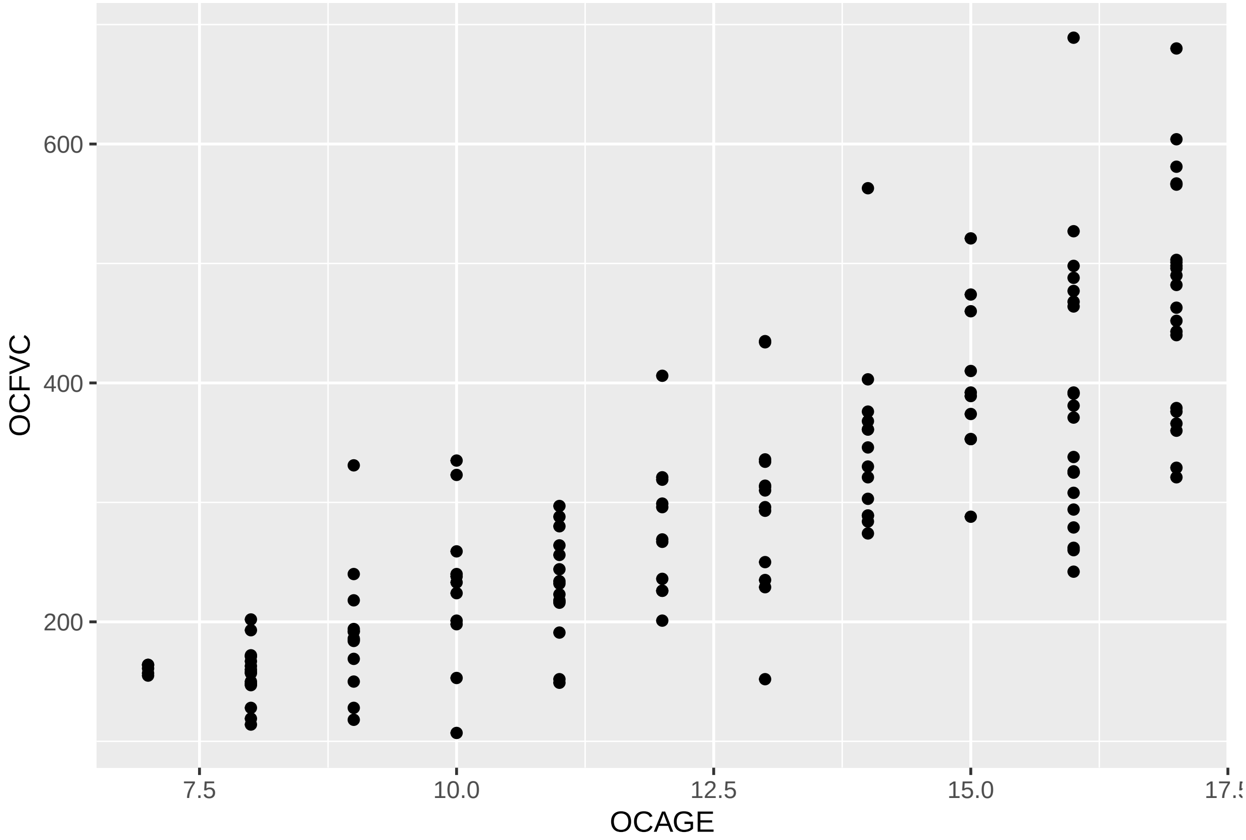
```
a1 <- group_by(lung, OCFVC, AREA.2)
summarise(a1)
```

```
## `summarise()` has grouped output by 'OCFVC'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 144 x 2
## # Groups:   OCFVC [128]
##    OCFVC AREA.2
##    <int> <fct>
##  1   107 Lancaster
##  2   114 Burbank
##  3   118 Glendora
##  4   119 Glendora
##  5   128 Lancaster
##  6   128 Long Beach
##  7   147 Glendora
##  8   148 Burbank
##  9   148 Glendora
## 10   149 Lancaster
## # ... with 134 more rows
```

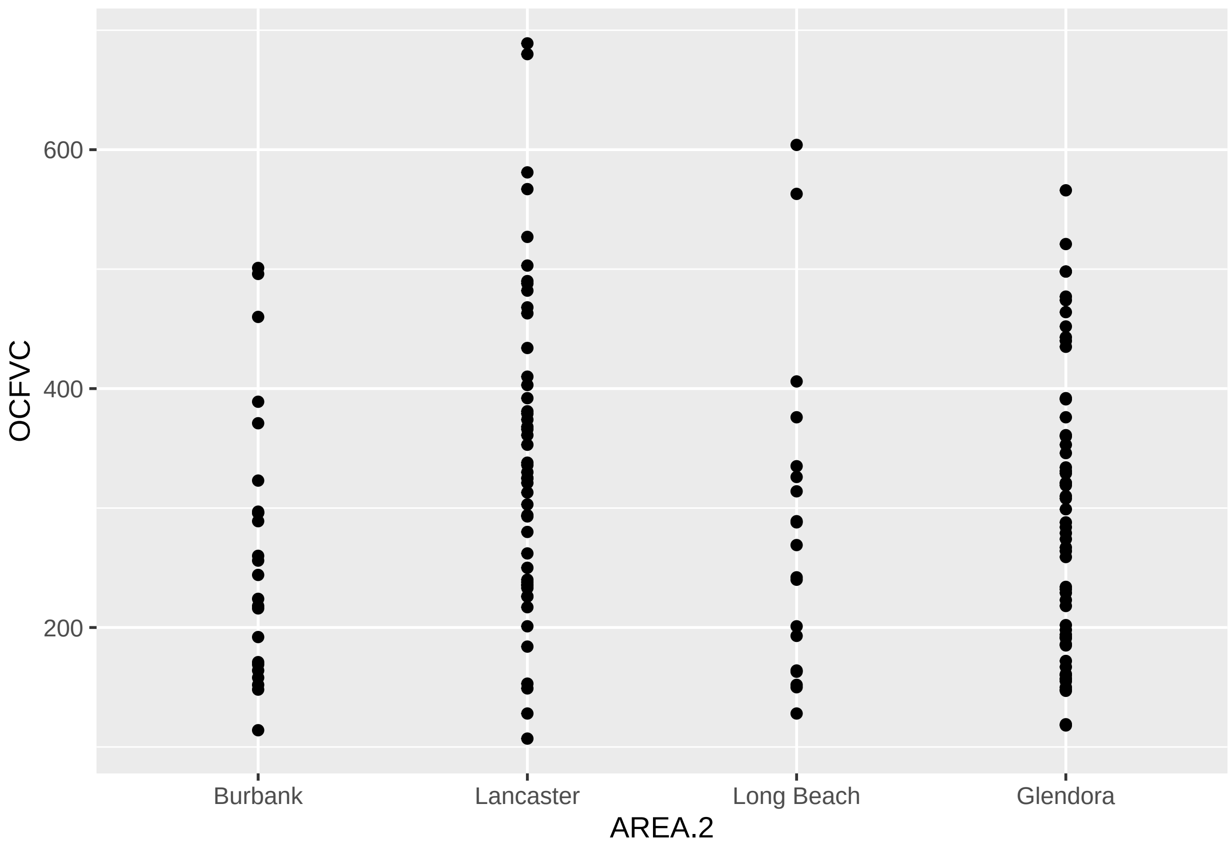
Above is the breath volume per the location and pollution of each area.

```
library(ggplot2)
ggplot(lung, aes(x=OCAGE, y=OCFVC)) + geom_point()
```



Above is the age of the oldest child relative to the breath volume of their lungs. We can see that as the children get older their breath volume increases.

```
ggplot(lung, aes(x=AREA.2, y=OCFVC)) + geom_point()
```



While their lung volume increases with age, the lung volume in each area stays relatively within the range of 0 to 400 and from there the frequency decreases.