Lab 1

Alyssa Andrichik

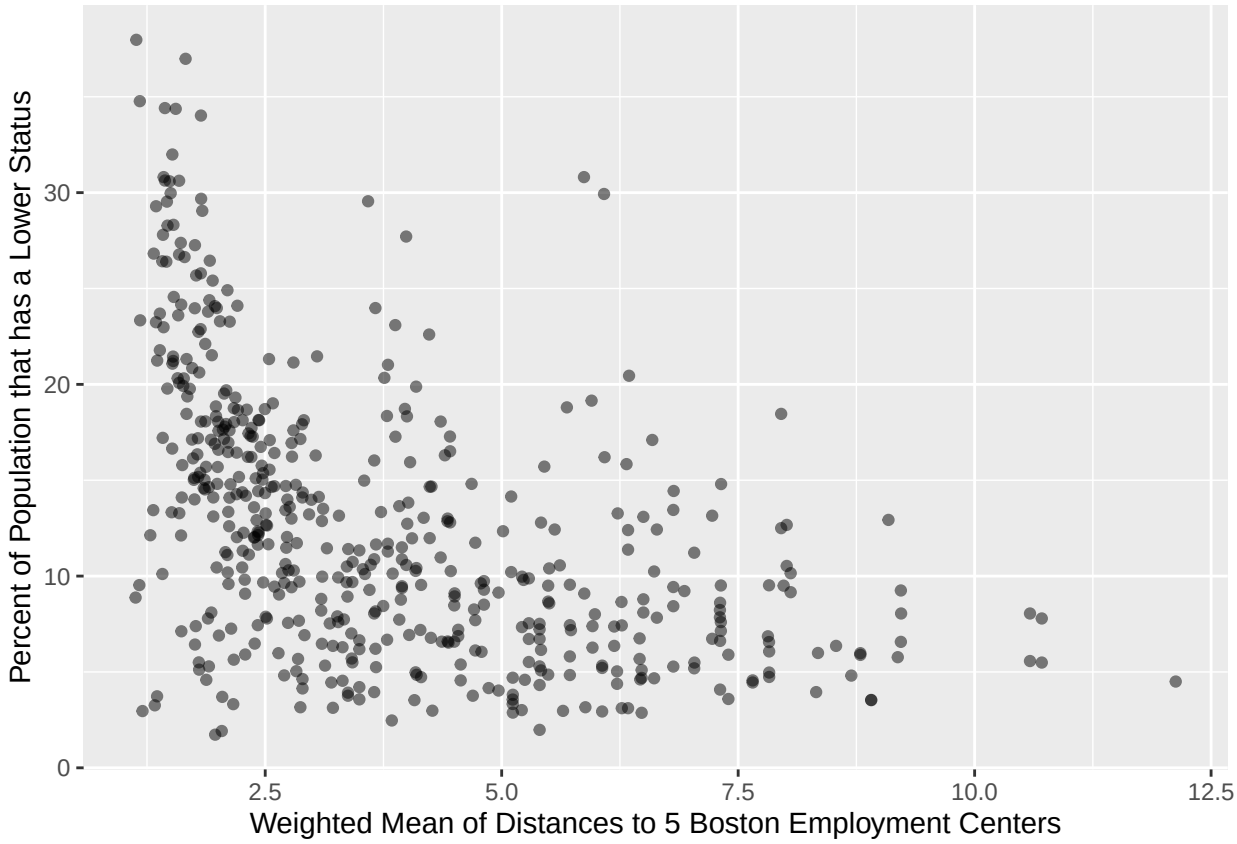
9/11/2019

1)How many rows are in this data set? How many columns? What do the rows and columns represent?

The Boston data frame has 506 rows and 14 columns. The rows represent each suburb of Boston and the values/observations of the variables. The columns represent different predictors/variables that aim to help figure out the average value of the houses in a specific Boston suburb. This data set is focused on creating an inference model, since the median value of a house in the specific suburb is a variable.

2)Make some (2-3) pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

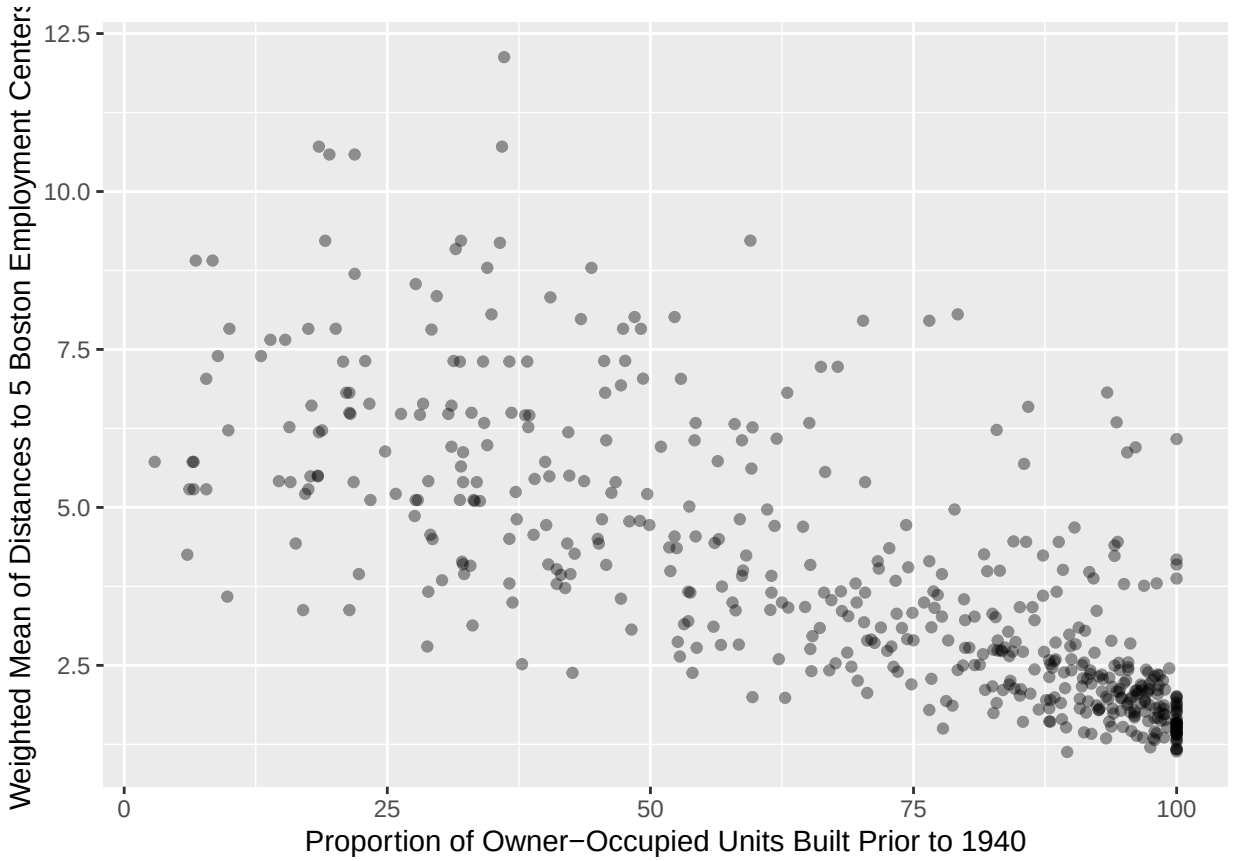
```
ggplot(data = Boston, mapping = aes(x = dis,
y = lstat)) +
geom_point(alpha = 0.5) +
  labs(x = "Weighted Mean of Distances to 5 Boston Employment Centers", y = "Percent of Population that
```



There is a negative exponential correlation between the weighted mean of distances from a suburb to 5 Boston

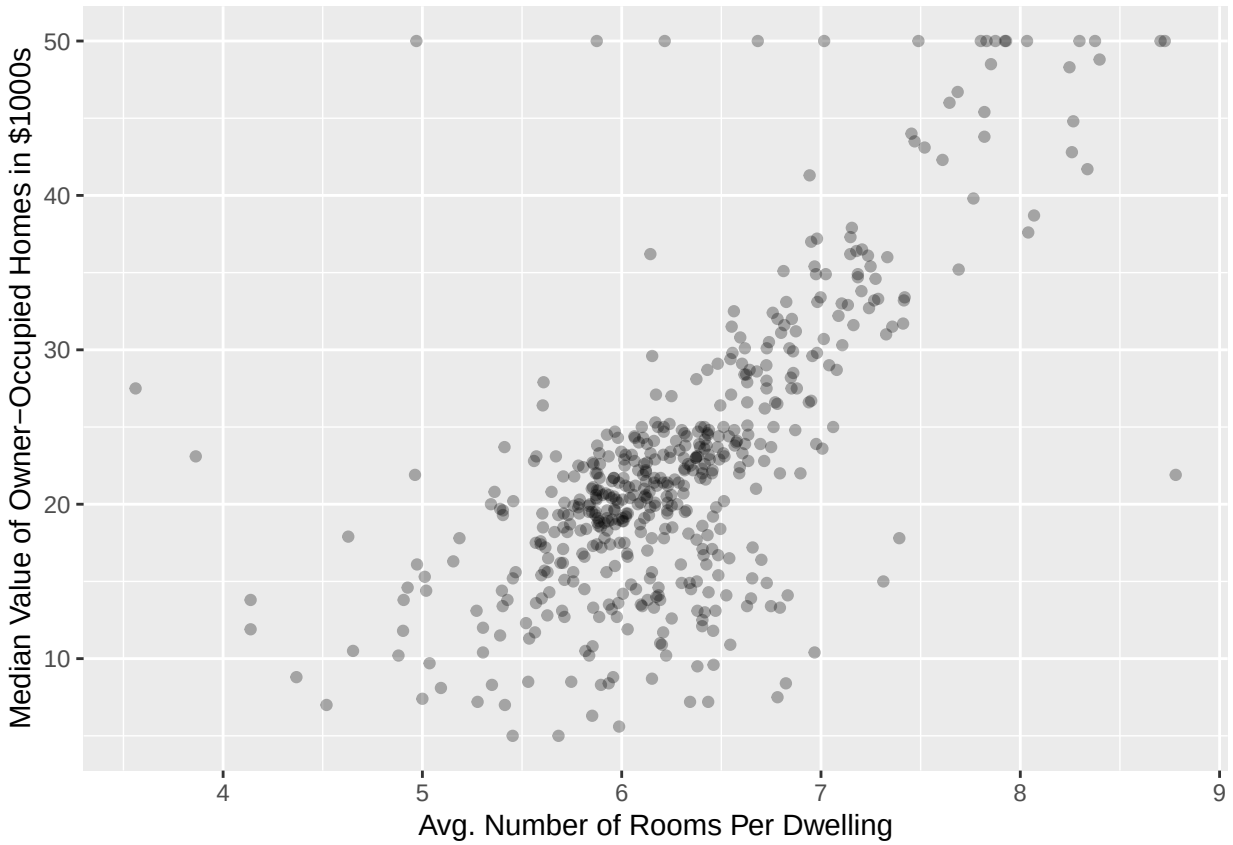
employment centers and the percent of population that has a lower status in a suburb. This means that, typically, the shorter the distance a suburb is from the weighted mean of distances to 5 Boston employment centers, a larger percent of the population of the suburb is considered to be of a lower status.

```
ggplot(data = Boston, mapping = aes(x = age,
y = dis)) +
geom_point(alpha = 0.4) +
  labs(x = "Proportion of Owner-Occupied Units Built Prior to 1940", y = "Weighted Mean of Distances to
```



There is a negative linear correlation between the proportion of owner-occupied units built prior to 1940 and the weighted mean of distances from a suburb to 5 Boston employment centers. This means that, typically, the suburbs with a larger proportion of the occupied units built prior to 1940 will be closer to the Boston employment centers. There is a smaller amount of suburbs with a greater amount of units built after 1940 in the data, but there is a distinct trend showing that newer units were built farther away from the employment centers.

```
ggplot(data = Boston, mapping = aes(x = rm,
y = medv)) +
geom_point(alpha = 0.3) +
  labs(x = "Avg. Number of Rooms Per Dwelling", y = "Median Value of Owner-Occupied Homes in $1000s")
```



There is seemingly positive linear correlation between the average number of rooms per dwelling in a suburb and the median value of owner-occupied homes in a suburb. Most suburbs' average number of rooms is around 6, so there are less observations of suburbs that typically have more or less than 6 rooms on average. There is a distinct positive correlation based on the observations, but more observations of suburbs that have an average number of rooms as more and less than 6 would be preferable to ensure that the positive correlation is accurate. The scatter-plot shows that the more rooms on average leads to a higher median value. This makes sense because the more rooms in a house typically means a house is bigger, which typically means that a house is more expensive.

3)Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

```
cor(Boston)
```

```
##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
```

```
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                 rm         age         dis          rad         tax
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304
## black    0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593
##            ptratio       black      lstat       medv
## crim     0.2899456 -0.38506394  0.4556215 -0.3883046
## zn      -0.3916785  0.17552032 -0.4129946  0.3604453
## indus    0.3832476 -0.35697654  0.6037997 -0.4837252
## chas    -0.1215152  0.04878848 -0.0539293  0.1752602
## nox      0.1889327 -0.38005064  0.5908789 -0.4273208
## rm      -0.3555015  0.12806864 -0.6138083  0.6953599
## age      0.2615150 -0.27353398  0.6023385 -0.3769546
## dis     -0.2324705  0.29151167 -0.4969958  0.2499287
## rad      0.4647412 -0.44441282  0.4886763 -0.3816262
## tax      0.4608530 -0.44180801  0.5439934 -0.4685359
## ptratio  1.0000000 -0.17738330  0.3740443 -0.5077867
## black   -0.1773833  1.00000000 -0.3660869  0.3334608
## lstat    0.3740443 -0.36608690  1.0000000 -0.7376627
## medv    -0.5077867  0.33346082 -0.7376627  1.0000000
```

Per capita crime rate (crim) has a high positive correlation with the index of accessibility to radial highways (rad). The correlation coefficient is 0.626. This means that there is a high crime rate in suburbs that have easy access to a highway leading to or from an urban center. Per capita crime rate (crim) also has a high positive correlation with the percent of lower status people (lstat) in a suburb. The correlation coefficient is 0.456. This means there is a higher crime rate in suburbs with a larger percent of lower status residents. Per capita crime rate (crim) has a high positive correlation with the full-value property-tax rate (tax) of a suburb. The correlation coefficient is 0.583. This means that there is a higher crime rate in suburbs with higher property-tax rates.

4)Are there any suburbs of Boston that appear to have particularly high crime rates? Tax rate? Pupil-teacher ratios? Comment on the range of each predictor.

The highest crime rates are in suburbs 381 (88.97620), 419 (73.53410), and 406 (67.92080). The range is 0.00632 to 88.97620.

The highest tax rates are in suburbs 489, 490, 491, 492, and 493 who's full-value property-tax rate per $10,000

is 711. The range is 187 to 711.

The highest pupil-teacher ratios are in suburbs 355 and 366 both at a ratio of 22 to 1. The range of this ratio is 12.6:1 to 22:1.

5)How many of the suburbs in this data set bound the Charles river?

```
Chas_values <- table(Boston$chas)
```

35 suburbs are bound to the Charles river, 471 are otherwise not.

6)What is the median pupil-teacher ratio among the towns in this data set?

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

The median pupil-teacher ratio among the towns is 19.05 pupils to 1 teacher.

7)If you want to build a model to predict the average value of a home based on the other variables, what is your output/response? What is your input?

The model's input would take into account the correlation between the median value of owner-occupied homes (medv) and the values/observations of the other variables. We want a flexible model that takes into account many parameters and a large sample size, so we would interpret the entire data set to understand trends between the medv variable and the others. The model would have to account for the correlation of each predictor to predict the output/response (the avg housing value). Based on how the predictors relate to medv, the model would interpret how each variable is applicable to medv. Then, based on how it is interpreted, my new model would apply the correlations to predict the average value of a home.