Final Project: Natural Amenities in the US

By: Cameron Deal, Vikram Meyer, and Zoe Luther

Introduction

Deciding where to live is a big decision with many big factors involved such as job location and family location. However, there are also smaller factors that play a role such as the availability of natural amenities including bodies of water, parks, mountains, and average temperature throughout the year. Given that small factors such as the availability of natural amenities can play a role in people's decisions, the goal of this project is to provide statistical estimates on the places to live with the best availability of natural amenities. We are also interested in the relationship between urban density and natural amenities, as well as urban density's mediating impact on the natural amenities scores of census divisions.

Here are the proposed hypotheses:

- H1: Counties in the Mountain and Pacific Census regions will have higher natural amenities scores given the frequency of national parks and other sites in these areas. The Midwest will have the lowest given the lack of associated natural beauty.
- H2: Urban counties will have higher natural amenities scores, as more people have chosen to live there given the environmental qualities of the region.

Data Exploration

```
#Load in Data and Libraries
library(ggplot2)
library(sjPlot)
library(tidyverse)
library(doBy)
library(maps)
library(stringr)
library(knitr)
library(data.table)
#install.packages("DT")
library(DT)
library(rgdal)
library(leaflet)
library(dplyr)
library(ggplot2)
library(broom)
library(usmap)
library(ggthemes)
library(ggdag)

amenities_data <- read.csv("amenities_scale.csv")

###Clean up variables and convert to what we need

#Census Divisions
amenities_data$cens_div <- factor(amenities_data$cens_div,
levels = c(1:9),
labels = c("New England", "Middle Atlantic", "East North Central", "West North Central", "South Atlantic", "East South Central", "West South Central", "Mountain", "Pacific"))


#States
amenities_data$state <- as.factor(amenities_data$state)

#Urban-Rural Code (varies from 0- Most Urban to 9- Most Rural)
```

The natural amenities scale data offers a measure of the geographical and physical qualities of a county that may make it desirable as a place to live. First, we loaded the data we needed from a csv file and used several commands to clean up the data.

Independent Variables

Census Divisions: The US Census Bureau divides the country into nine regions that are geographically contiguous. We labelled the census division variable and converted it to a factor variable from a numeric variable.

Rural-Urban Continuum Code: This measure, which ranges from 0 (most urban) to 9 (rural) captures the population density of a county by measuring if the county is in a metropolitan area and, if so, the population of the metropolitan area.

Auxiliary Variables: We converted the state identifiers to a factor variable from a string variable, allowing for easier comparisons. Our other variables of interest were primarily numeric, so we did not alter them.

Dependent Variables

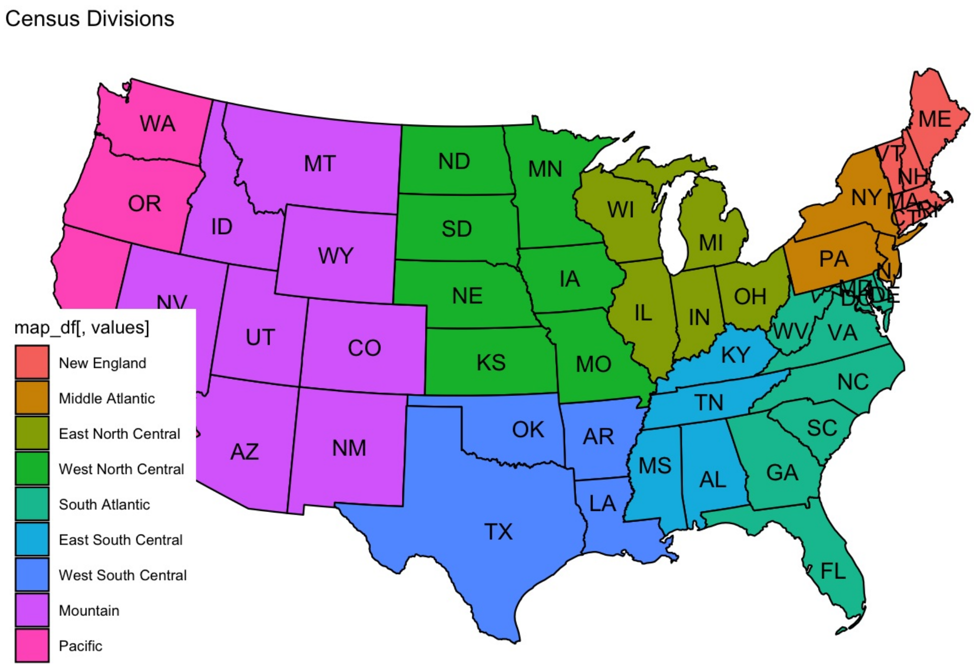
Natural Amenities Scale: The scale is constructed through six measures: warm winter, winter sun, temperate summer, low summer humidity, topographic variation, and water area. It is standardized around 0, with a standard deviation of 1 and ranges from -6.40 to 11.17 in this dataset.

Visualizations

First, we use the graphic capabilities of R to visualize some of the contributing factors to the natural amenities score.

To contextualize the census divisions in question, the following map diagrams the geographic breakdown of each of the 9 regions of the continental US. The dataset we used did not include data from Alaska and Hawaii so those states are omitted from our graphical interpretations.

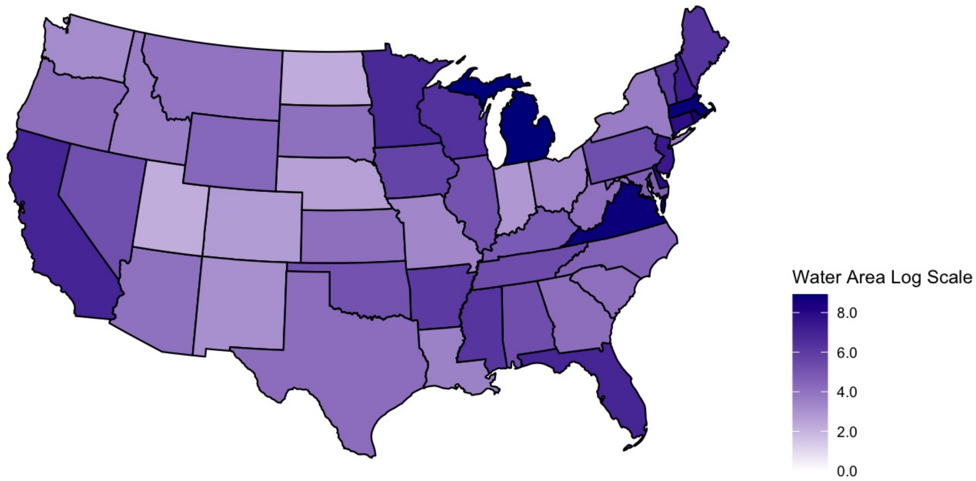
```
amenities_data = amenities_data %>%
  mutate(fips = str_pad(as.character(fips), 5, pad="0"))

cens_div_data = map_with_data(amenities_data, values = "cens_div",na = NA)
cens_div_data = dplyr::select(cens_div_data, state, cens_div)

plot_usmap(
regions = c("states"),
exclude= c("AK","HI"),
data = cens_div_data,
values = "cens_div",
theme = theme_map(),
labels = TRUE,
label_color = "black"
) + labs(title = "Census Divisions")
```



The map below depicts each state's quantity of water area (logarithmic scaling) which is one of several factors that compose the overall natural amenities score. As one may anticipate, the coastal states and those with proximity to the Great Lakes exhibit significantly larger `water_area` values, thus aligning with the expectation that coastal areas would exhibit higher natural amenities scores.

```
county_level_water_data = map_with_data(amenities_data, values = "water_area_log", na=NA)
county_level_water_data_WATER = dplyr::select(county_level_water_data, state, water_area_log)

plot_usmap(
regions = c("states"),
exclude= c("AK","HI"),
data = county_level_water_data_WATER,
values = "water_area_log",
theme = theme_map(),
labels = FALSE,
label_color = "grey"
) +
  scale_fill_continuous(
    low = "white", high = "blue4", name = "Water Area Log Scale", label = scales::comma
  ) + theme(legend.position = "right")  + labs(title = "Water Area")
```

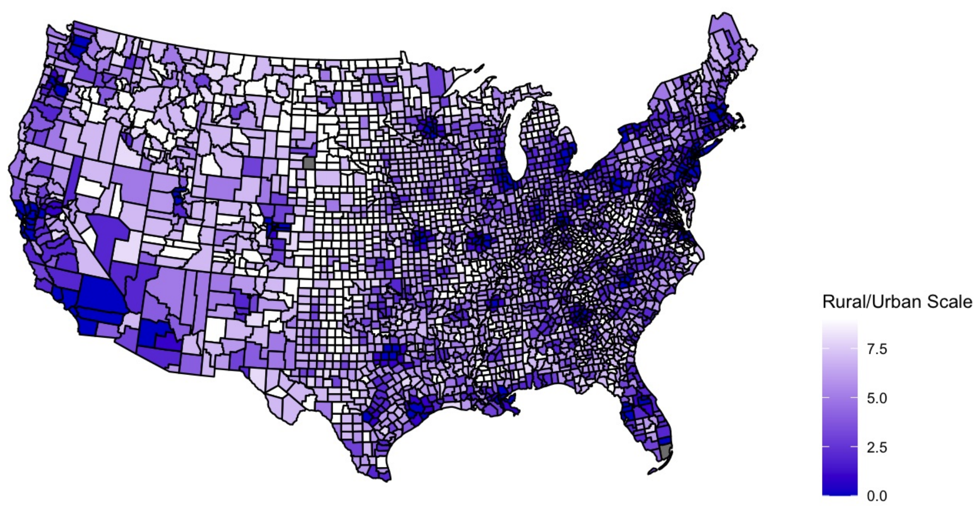
Water Area



The following map displays the county-level data on the rural-ness of the area (a higher score on the Rural/Urban Scale indicates a more rural region). The darker (more urban) regions in the graph unsurprisingly align with more coastal regions and are skewed toward the Eastern United States. When evaluating the relationship between the rural-ness of a region and its natural amenities score, the relationship is more ambiguous than a simple causal link. We can observe an association between certain coastal regions and higher urban density while also observing that the natural beauty of more rural Mid-Western areas is associated with a lower urban density.

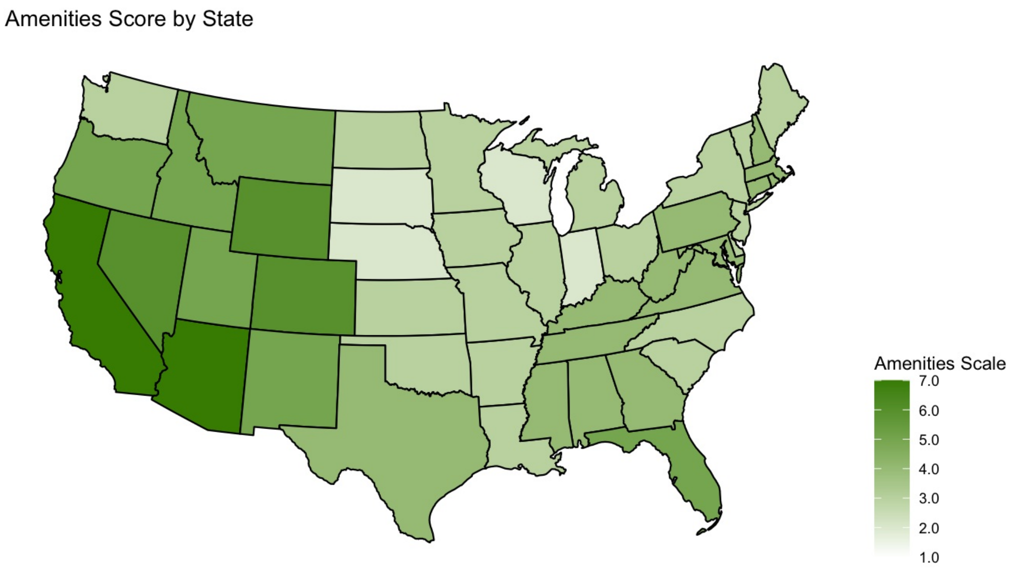
```
county_level_urban_data = map_with_data(amenities_data, values = "rural_urban", na=NA)
county_level_urban_data_RURAL = dplyr::select(county_level_urban_data, fips, rural_urban)

plot_usmap(
regions = c("states"),
exclude= c("AK","HI"),
data = county_level_urban_data_RURAL,
values = "rural_urban",
theme = theme_map(),
labels = FALSE,
label_color = "grey"
) +
  scale_fill_continuous(
    low = "mediumblue", high = "white", name = "Rural/Urban Scale", label = scales::comma
  ) + theme(legend.position = "right")  + labs(title = "Rural/Urban Score by County")
```

Rural/Urban Score by County



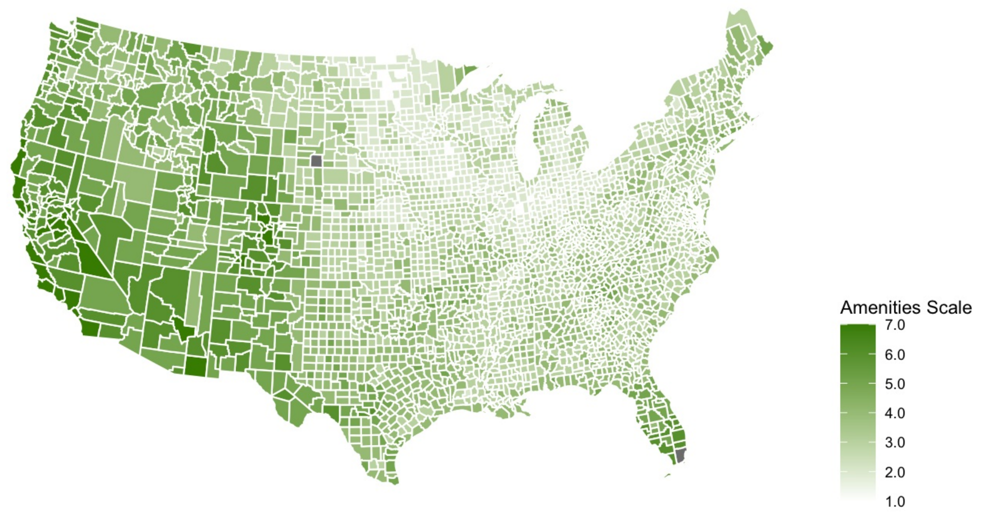
The following two graphs depict the composite natural amenities score by state and by county. As anticipated, Western and coastal areas display significantly higher scores while the Eastern and Central US exhibit a lack of natural amenities according to this data set.

```
state_level_data = map_with_data(amenities_data, values = "scale_seven", na=NA)
state_level_data_SCALE_SEVEN = dplyr::select(state_level_data, state, scale_seven)

plot_usmap(
regions = c("states"),
exclude= c("AK","HI"),
data = state_level_data_SCALE_SEVEN,
values = "scale_seven",
theme = theme_map(),
labels = FALSE,
label_color = "black"
) +
  scale_fill_continuous(
    low = "white", high = "chartreuse4", name = "Amenities Scale", label = scales::comma
  ) + theme(legend.position = "right") + labs(title = "Amenities Score by State")
```



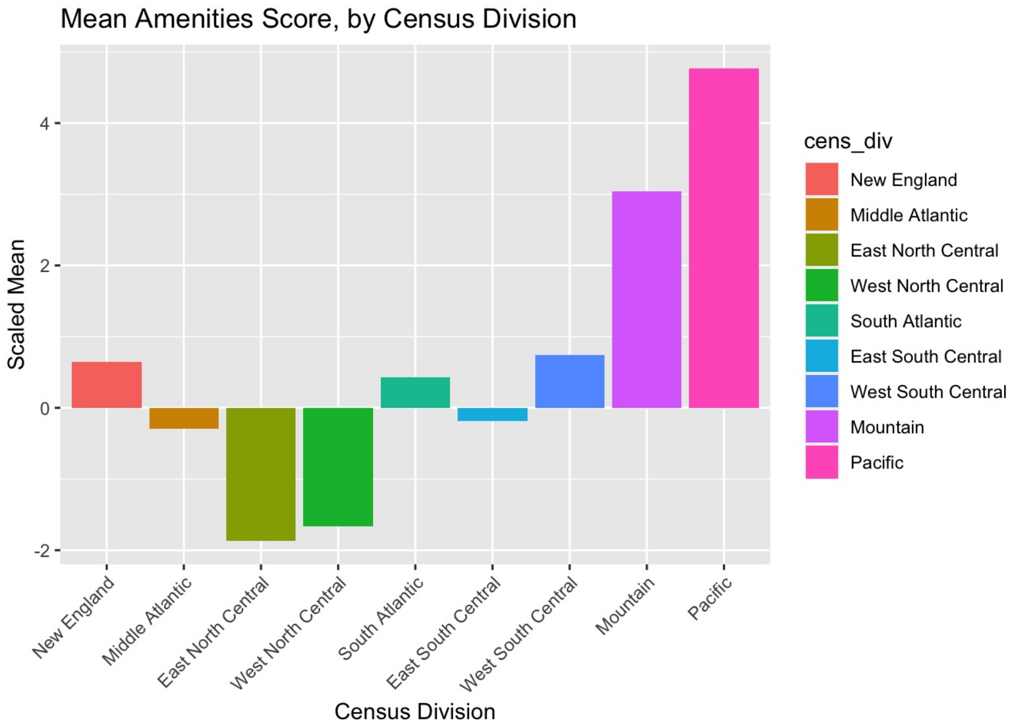

```
county_level_data = map_with_data(amenities_data, values = "scale_seven",na = NA)
county_level_data_SCALE_SEVEN = dplyr::select(county_level_data, fips, scale_seven)

plot_usmap(
regions = c("states"),
exclude= c("AK","HI"),
data = county_level_data_SCALE_SEVEN,
values = "scale_seven",
theme = theme_map(),
labels = FALSE,
label_color = "grey",
color = "white"
) +
  scale_fill_continuous(
    low = "white", high = "chartreuse4", name = "Amenities Scale", label = scales::comma
  ) + theme(legend.position = "right")  + labs(title = "Amenities Score by County")
```

Amenities Score by County



Analysis & Interpretation

```
# Unadjusted Means for each region
division_means <- summaryBy(scale  ~ cens_div, data=amenities_data , FUN=c(mean), na.rm=TRUE)

ggplot(data=division_means, aes(x=cens_div, y=scale.mean, fill=cens_div)) + geom_bar(stat = "identity") + theme(a
xis.text.x = element_text(angle = 45, hjust = 1)) + xlab("Census Division") + ylab("Scaled Mean") + ggtitle("Mean
Amenities Score, by Census Division")
```



First, we constructed a set of mean scores for each census division and displayed the findings visually in a bar graph. We see that, on average, the highest scoring census divisions are Mountain and Pacific, and the lowest scoring census divisions are the East North Central and West North Central. These findings suggest that the Western United States is the region of the country with the highest natural amenities, while the Midwest has the lowest natural amenities. We explore these patterns further through a regression analysis.

```
model1 <- lm(scale ~ cens_div + rural_urban, data=amenities_data)
tab_model(model1, show.se=TRUE, digits = 3)
```

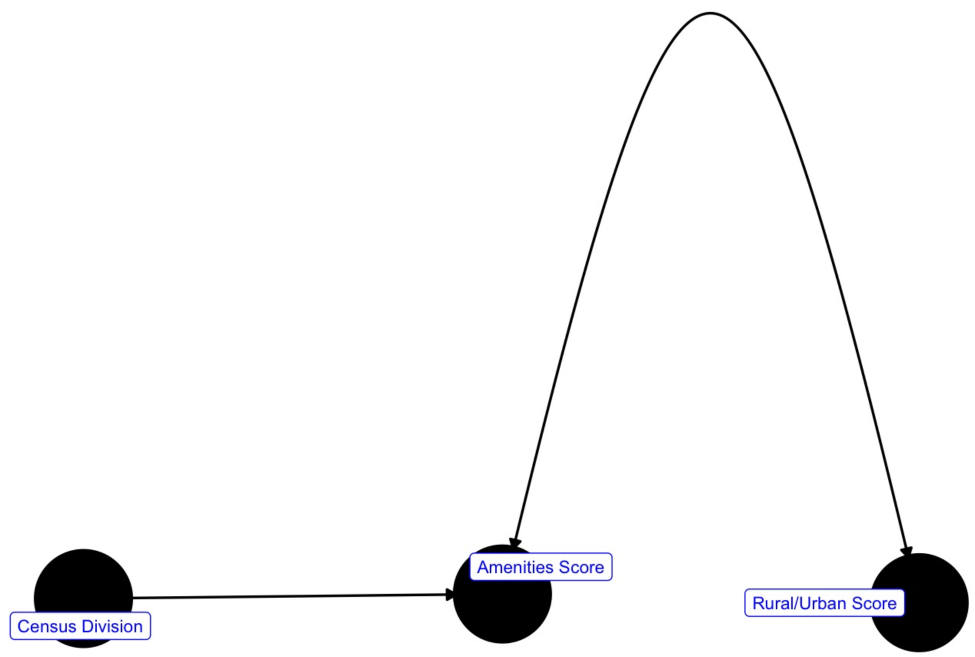
	scale			
Predictors	*Estimates*	*std. Error*	*CI*	*p*
(Intercept)	0.800	0.187	0.434 – 1.167	**<0.001**

cens div [Middle Atlantic]	-0.982	0.219	-1.411 – -0.553	**<0.001**
cens div [East North Central]	-2.484	0.195	-2.867 – -2.101	**<0.001**
cens div [West North Central]	-2.209	0.193	-2.588 – -1.830	**<0.001**
cens div [South Atlantic]	-0.190	0.192	-0.566 – 0.186	0.321
cens div [East South Central]	-0.762	0.199	-1.152 – -0.373	**<0.001**
cens div [West South Central]	0.154	0.195	-0.228 – 0.536	0.430
cens div [Mountain]	2.490	0.204	2.090 – 2.889	**<0.001**
cens div [Pacific]	4.132	0.223	3.695 – 4.569	**<0.001**
rural urban	-0.037	0.011	-0.058 – -0.017	**<0.001**
Observations	3111			
R^2 / R^2 adjusted	0.576 / 0.575			

Model 1 examines the natural amenities scale as the dependent variable and examines the role of census division while adjsuting for urban density through the inclusion of the urban-rural scale as a control variable. New England is the omitted reference group. While controlling for urban density, four regions scored significantly lower than New England on the natural amenities scale: Middle Atlantic (β=-0.98; 95% CI=(-1.41,-0.55)), East North Central (β=-2.48; 95% CI=(-2.88,-2.10)), West North Central (β=-2.21; 95% CI=(-2.59,-1.83)), and East South Central (β=-0.76; 95% CI=(-1.15,-0.37)). Additionally, there were two regions that scored significantly better than New England while controlling for urban density: Mountain (β=2.49; 95% CI=(2.09,2.89)), and Pacific (β=4.13; 95% CI=(3.70,4.57)). These findings confirm the patterns suggested by the descriptive statistics above; the Western United States enjoys considerably higher natural amenities than the Midwest and more central regions of the country.

Additionally, we were interested in the association between rural-urban density and natural amenities score. Using a similar model, we examined that relationship while controlling for regional differences through the inclusion of the census division variable as a control. We found that a one-point increase on the rural-urban continuum code (range 0-9) was associated with a 0.037-point decrease in the natural amenities scale (95% CI=(-0.058,-0.017), P-Value<0.001). This suggests that more rural counties tended to have lower natural amenities scores. To be clear, this is not evidence of a causal relation- it is very plausible that natural amenities influence where people live, and in turn urban density, violating the exogeneity necessary to establish causality and raising a potential issue of reverse causality. These relationships are seen in the below Directed Acyclic Graph. However, as a descriptive observation, this association establishes that more rural counties tend to have lower natural amenities in the United States, even when adjusting for regional differences.

```
theme_set(theme_dag())
dagify(amenity_score  ~~  urban_rural_score,
       amenity_score ~ census_division, labels = c("amenity_score" = "Amenities Score",
                  "census_division" = "Census Division",
                  "urban_rural_score" = "Rural/Urban Score")) %>%
  ggdag(text = FALSE, use_labels = "label", text_size = 3, node_size= 22, label_col = "blue")
```



```
model1 <- lm(scale ~ cens_div, data=amenities_data)
coeffs <- summary(model1)$coefficients
means_from_lm <-coeffs[1] + c(0, coeffs[2:length(levels(amenities_data$cens_div))])

# Division means, obtained directy, do coincide with those obtained from regression coeffs (as they should)

means_from_lm
```

```
## [1]  0.6440299 -0.2967333 -1.8667816 -1.6663710  0.4242978 -0.1864560  0.7410638
## [8]  3.0419217  4.7718797
```

```
model2 <- lm(scale ~ cens_div + rural_urban, data=amenities_data)
coeffs <- summary(model2)$coefficients
means_from_lm2 <-coeffs[1] + c(0, coeffs[2:length(levels(amenities_data$cens_div))])

# Take a look at what division means controlled for urban_rural are
means_from_lm2
```

```
## [1]  0.80028625 -0.18207090 -1.68395777 -1.40868202  0.60991885  0.03798622
## [7]  0.95394774  3.28994598  4.93212223
```

```
# Comparing means_from_lm & means_from_lm2 - no "exact" relation occurs
means_from_lm / means_from_lm2
```

```
## [1]  0.8047494  1.6297681  1.1085679  1.1829291  0.6956627 -4.9085176  0.7768390
## [8]  0.9246114  0.9675104
```

```
means_from_lm2 / means_from_lm
```

```
## [1]  1.2426229  0.6135842  0.9020647  0.8453592  1.4374782 -0.2037275  1.2872680
## [8]  1.0815354  1.0335806
```

```
means_from_lm - means_from_lm2
```

```
## [1] -0.1562564 -0.1146624 -0.1828238 -0.2576889 -0.1856210 -0.2244423 -0.2128839
## [8] -0.2480243 -0.1602425
```

```
# One "strange" thing is that means_from_lm2 are all larger than means_from_lm2 -
# that shouldn't be happening. Let's do a check: weighted (by count) mean of means should equal to overall mean

division_counts <- summaryBy(scale  ~ cens_div, data=amenities_data , FUN=length)$scale.length
division_counts
```

```
## [1]  67 150 435 620 591 364 470 281 133
```

```
mean(amenities_data$scale)
```

```
## [1] 0.05595307
```

```
sum(means_from_lm*division_counts) / sum(division_counts) # Coincide
```

```
## [1] 0.05595307
```

```
sum(means_from_lm2*division_counts) / sum(division_counts) # Differs
```

```
## [1] 0.2647045
```

```
# So, the overall mean, when computed from means_from_lm2 differs from the original,
# while it shouldn't. The reason for that is in lm2 there's a coefficient for
# urban_rural (cur), so we need to "shift the mean back" by cur * mean(urban_rural)

summary(model2)$coefficients
```

```
##                               Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)                 0.80028625 0.18688104   4.2823299 1.905595e-05
## cens_divMiddle Atlantic    -0.98235715 0.21877589  -4.4902440 7.373573e-06
## cens_divEast North Central -2.48424402 0.19526232 -12.7225980 3.498771e-36
## cens_divWest North Central -2.20896827 0.19331417 -11.4268306 1.193747e-29
## cens_divSouth Atlantic     -0.19036740 0.19182942  -0.9923786 3.210903e-01
## cens_divEast South Central -0.76230003 0.19857218  -3.8389064 1.260515e-04
## cens_divWest South Central  0.15366149 0.19480017   0.7888160 4.302799e-01
## cens_divMountain            2.48965972 0.20377282  12.2178203 1.457453e-33
## cens_divPacific             4.13183598 0.22273435  18.5505107 6.208736e-73
## rural_urban                -0.03738992 0.01052399  -3.5528275 3.867940e-04
```

```
cur = summary(model2)$coefficients["rural_urban",1]
cur * mean(amenities_data$rural_urban)
```

```
## [1] -0.2087514
```

```
means_from_lm2 = means_from_lm2 + cur * mean(amenities_data$rural_urban)
sum(means_from_lm2*division_counts) / sum(division_counts) # Now works fine
```

```
## [1] 0.05595307
```

```
# All in all, we need to make this amend for computing means straight away,
# that is, the formula for means from a linear model with control should look like

means_from_lm2 <- coeffs[1] +
      c(0, coeffs[2:length(levels(amenities_data$cens_div))]) +
      summary(model2)$coefficients["rural_urban",1] * mean(amenities_data$rural_urban)
sum(means_from_lm2*division_counts) / sum(division_counts) # Now works fine
```

```
## [1] 0.05595307
```

```
# Now let's compare them once again
means_from_lm / means_from_lm2
```

```
## [1] 1.0887437 0.7592539 0.9863014 1.0302563 1.0576576 1.0918856 0.9944545
## [8] 0.9872540 1.0102700
```

```
means_from_lm2 / means_from_lm
```

```
## [1] 0.9184898 1.3170826 1.0138889 0.9706323 0.9454856 0.9158468 1.0055765
## [8] 1.0129105 0.9898344
```

```
means_from_lm - means_from_lm2
```

```
## [1]  0.052494997  0.094088963  0.025927560 -0.048937550  0.023130351
## [6] -0.015690869 -0.004132513 -0.039272870  0.048508864
```

Once controlling for the urban-rural code, the adjusted means for each census division have changed to varying degrees. These changes are expressed above through the ratios and differences between the adjusted means and the original means. The census division mean for the Middle Atlantic region increased significantly when controlling for the urban-rural continuum (Mean Ratio= 1.317), indicating that this division possesses higher relative natural resource scores once adjusting for urban density. Several others, including New England (MR=0.918), South Atlantic (MR=0.945), and East South Central (MR=0.916), had lower natural resource scores once controlling for urban density.