USDA Data Science Training Program  
Intermediate Assignment 3: Exploratory Data Analysis

**Data:** USDA-ERS data on food environment factors (a subset of the Food Environment Atlas), along with county-level food security data from Feeding America. This assignment uses the variables listed below.

* State - Name of the state
* County - Name of the county
* CHILDPOVRATE15 - Child poverty rate, 2015
* PCT\_LACCESS\_HHNV15 - Percent of households with no car & low access to store, 2015
* METRO13 - Metro/nonmetro classification, 2013

For reference, a full list of the variables in this dataset is included at the end of the assignment.

**DataCamp:** The following DataCamp courses correspond to this exercise:

* R: Exploratory Data Analysis in R, Introduction to Statistics in R, Reporting with R Markdown (bonus question)
* Python: Exploratory Data Analysis in Python, Introduction to Statistics in Python, Building Dashboards with Dash and Plotly (bonus question)

**Assignment:**

*Part A: Load, inspect, and prep the data*

1. Save EDA\_assignment\_dataset.csv to your computer and import the data into a dataframe called food\_env.
2. Print the summary statistics for all columns in food\_env. Do any variables have missing data?
3. Create a new dataframe named food\_env\_full that excludes rows with any missing data. How many rows were dropped?
4. Convert the METRO13 column to data type “logical.”

## FIPS State County FMRKT\_SNAP18   
## Min. : 1001 Length:3142 Length:3142 Min. : 0.0000   
## 1st Qu.:18178 Class :character Class :character 1st Qu.: 0.0000   
## Median :29176 Mode :character Mode :character Median : 0.0000   
## Mean :30384 Mean : 0.9143   
## 3rd Qu.:45081 3rd Qu.: 1.0000   
## Max. :56045 Max. :53.0000   
## NA's :2   
## METRO13 PCT\_LACCESS\_HHNV15 GROC16 Pop2020   
## Min. :0.0000 Min. : 0.000 Min. : 0.00 Min. : 64   
## 1st Qu.:0.0000 1st Qu.: 1.683 1st Qu.: 2.00 1st Qu.: 10837   
## Median :0.0000 Median : 2.672 Median : 5.00 Median : 25749   
## Mean :0.3713 Mean : 3.267 Mean : 20.82 Mean : 105520   
## 3rd Qu.:1.0000 3rd Qu.: 4.034 3rd Qu.: 12.00 3rd Qu.: 68016   
## Max. :1.0000 Max. :53.515 Max. :2495.00 Max. :10014009   
## NA's :2 NA's :4 NA's :2 NA's :1   
## POVRATE15 SNAPS17 CHILDPOVRATE15 food\_insecurity\_2016  
## Min. : 3.40 Min. : 0.083 Min. : 3.30 Min. :0.037   
## 1st Qu.:11.50 1st Qu.: 10.417 1st Qu.:16.25 1st Qu.:0.110   
## Median :15.20 Median : 24.042 Median :22.30 Median :0.132   
## Mean :16.26 Mean : 73.536 Mean :23.24 Mean :0.137   
## 3rd Qu.:19.70 3rd Qu.: 58.188 3rd Qu.:29.10 3rd Qu.:0.157   
## Max. :47.40 Max. :6112.833 Max. :61.60 Max. :0.361   
## NA's :3 NA's :26 NA's :3

## [1] "The number of dropped rows is 29"

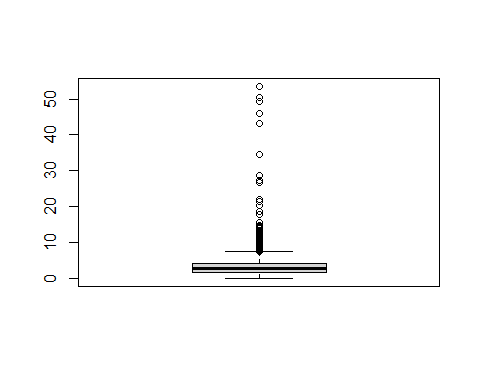
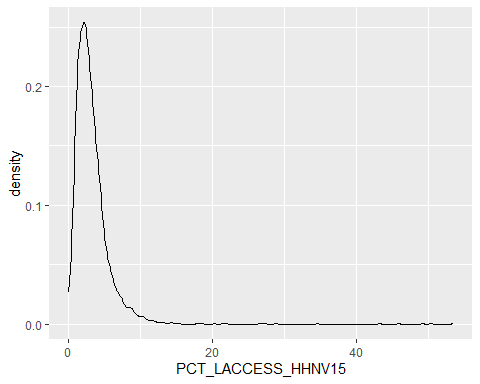
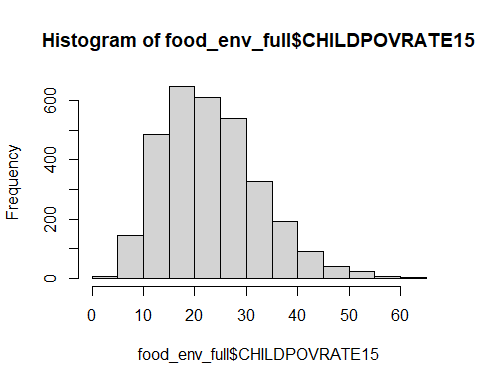
## [1] "The data type of METRO13 is logical"

*Part B: Explore the data*

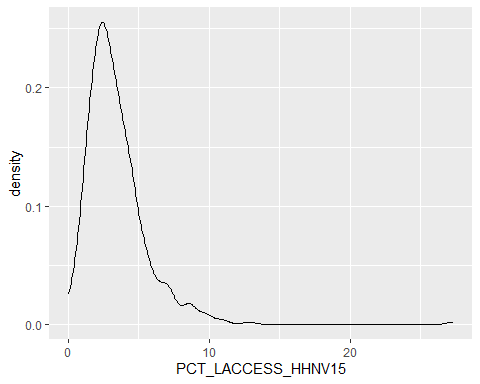
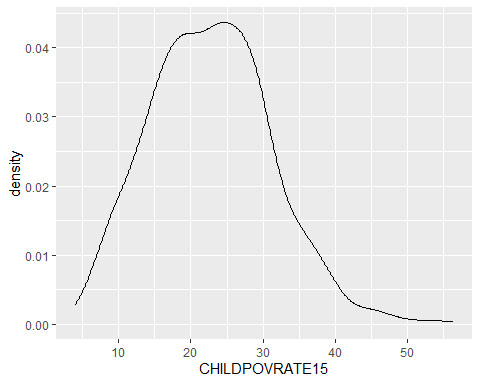
1. What are the minimum, maximum, mean, median, and standard deviation of CHILDPOVRATE15? Plot a histogram of child poverty rates to view the distribution.
2. Create a density plot to examine the distribution of PCT\_LACCESS\_HHNV15. Does it look like the data are skewed? To investigate further, create a boxplot of the PCT\_LACCESS\_HHNV15 data.
3. Filter food\_env\_full to just counties in Virginia, Illinois, Michigan, Arkansas, and Ohio and save the smaller dataset as food\_env\_5\_states.
4. What are the minimum, maximum, mean, median, and standard deviation values of CHILDPOVRATE15 and PCT\_LACCESS\_HHNV15 in each state?
5. Create faceted histograms or density plots of CHILDPOVRATE15 and PCT\_LACCESS\_HHNV15 by state. What do you notice about the distributions?

## [1] "Child poverty rate statistics"

## [1] "Minimum: 3.3"  
## [1] "Maximum: 61.6"  
## [1] "Mean: 23.2802762608416"  
## [1] "Median: 22.4"  
## [1] "Standard Deviation: 9.36221328449228"  
## [1] "Standard Deviation: 9.36221328449228"



## # A tibble: 5 × 11  
## State minimumPOV maximumPOV meanPOV medianPOV standard\_deviationPOV minimumACC  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 AR 15.3 56.4 31.2 30.2 7.57 1.25   
## 2 IL 5.5 52.2 20.2 19.6 6.85 0.444   
## 3 MI 7.8 45.2 23.5 24.1 7.14 1.10   
## 4 OH 4.7 33.3 20.4 20.4 6.75 1.23   
## 5 VA 4 43 21.0 21.3 8.80 0.0208  
## # ℹ 4 more variables: maximumACC <dbl>, meanACC <dbl>, medianACC <dbl>,  
## # standard\_deviationACC <dbl>

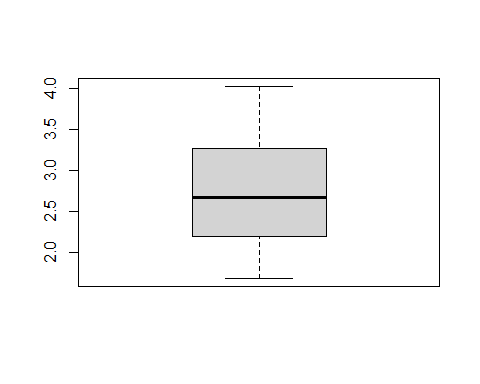


## [1] "Poverty rate is normally distributed while access is right skewed"

*Part C: Preliminary analysis*

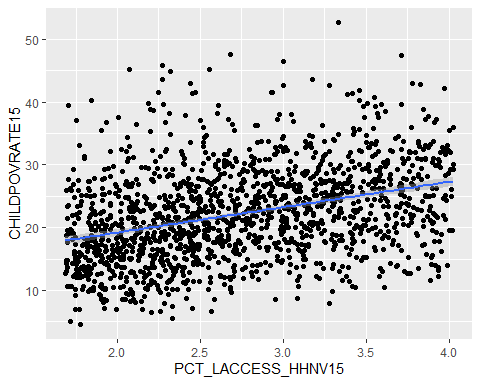
1. Using food\_env\_full, compute the interquartile range of the PCT\_LACCESS\_HHNV15 column to identify outliers. Remove rows with outliers from the data then create a new boxplot with the filtered dataframe. How does this boxplot compare to the one you created in question 6?
2. Make a scatterplot with PCT\_LACCESS\_HHNV15 (outliers removed) on the x-axis and CHILDPOVRATE15 on the y-axis. Add a straight line that shows the linear relationship between the two variables to your plot.
3. Compute the correlation between PCT\_LACCESS\_HHNV15 and CHILDPOVRATE15. Do you think the relationship is strong? Are there other factors that could affect both the child poverty rate and a household’s access to a grocery store in a county?

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.687 2.672 3.252 4.026 53.515



## [1] "The data is no longer skewed in comparison to the boxplot created in question 6."

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



##   
## Call:  
## lm(formula = CHILDPOVRATE15 ~ PCT\_LACCESS\_HHNV15, data = OutliersRemoved)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.4586 -4.8231 -0.5415 4.2230 28.1251   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.1365 0.7666 14.53 <2e-16 \*\*\*  
## PCT\_LACCESS\_HHNV15 4.0323 0.2727 14.79 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.947 on 1554 degrees of freedom  
## Multiple R-squared: 0.1234, Adjusted R-squared: 0.1228   
## F-statistic: 218.7 on 1 and 1554 DF, p-value: < 2.2e-16

## [1] "The relationship is significant but not strong."

## [1] "Parental education level and ethnic classification are likely better predictors of child poverty rate."

**Bonus:** Create a report in R Markdown or Dash using at least one of the plots you made, a table showing summary statistics for either CHILDPOVRATE15 or PCT\_LACCESS\_HHNV15, and text explaining your analysis.

**Deliverables:**

* Your code (the .R, .Rmd, .py, or .ipynb file).
* If you’ve chosen to write your responses or to complete the bonus question in an R Markdown file, the knitted document with your responses.
* These deliverables will be submitted through GitHub by the end of the program.

**Data Dictionary:**

* FIPS - Numeric code that identifies the geographic area
* State - Name of the state
* County - Name of the county
* Pop2020 - Population of the county, 2020
* POVRATE15 - Poverty rate, 2015
* SNAPS17 - Count of SNAP-authorized stores in the county, 2017
* CHILDPOVRATE15 - Child poverty rate, 2015
* PCT\_LACCESS\_HHNV15 - Percent of households with no car & low access to store, 2015
* GROC16 - Count of grocery stores in the county, 2016
* FMRKT\_SNAP18 - Count of Farmers’ markets that report accepting SNAP, 2018
* METRO13 - Metro/nonmetro classification, 2013
* food\_insecurity\_2016 - Food insecurity rate, 2016