HW2-Darshan Patel-3:30-5:30PM

Darshan Patel

9/28/2018

Deprivation indices are used to measure levels of deprivation in populations which go beyond simply measuring income (as in poverty measures). Initially developed in the UK, the most common are the Townsend, Carstairs, Jarman, and the Index of Multiple Deprivation. These indices have been shown to be good proxies for indicating health-related problems in local communities.

Goal: In this assignment, you will compute the Townsend Material Deprivation Index for census tracts in New York County (i.e., Manhattan) using American Community Survey (ACS) data administered by the US Census Bureau.

The Townsend index is constructed from four variables: % unemployment among people 16 and over (unemp), % of overcrowded households defined as homes with more than one person per room (oc), % of housing which is rented (rent) and % of households without a vehicle (car). A higher percentage of any of these variables indicates more deprivation in a geographic area (Note: do not convert the percentage variables into decimal form). Let geographic region be denoted by the subscript i and variable by j. To compute the index, follow the four steps described next. (a) Transform the variables to reduce skewness

$$\begin{aligned} t_{i,\text{unemp}} &= \log(x_{i,\text{unemp}} + 1) \\ t_{i,\text{oc}} &= \log(x_{i,\text{oc}} + 1) \\ t_{i,\text{rent}} &= \log(x_{i,\text{rent}} + 1) \\ t_{i,\text{car}} &= \sqrt{x_{i,\text{car}}} \end{aligned}$$

(b) Compute geographic region mean $\bar{t}_j$ and standard deviation s_j for each variable j. (c) Standardize variables for each region i and variable j:

$$z_{ij} = \frac{t_{ij} - \bar{t}_{ij}}{s_j}$$

(d) Compute sum of standardized variables for each region i:

$$\text{Townsend}_i = \sum_{j=1}^{4} z_{ij}$$

Have one Townsend index value for each region i. These values are on a relative scale; that is, the actual number is meaningless, just the ranking is important. A negative score indicates a **less** deprived region; a score of 0 indicates roughly the average level of deprivation; a positive score indicates a **more** deprived region.

```r
# Import packages
library(tidyverse)
```

```
## ── Attaching packages ──────────────────────────────────────────────────────
──────── tidyverse 1.2.1 ──

## ✔ggplot2 2.2.1     ✔purrr   0.2.5
## ✔tibble  1.4.2     ✔dplyr   0.7.6
## ✔tidyr   0.8.1     ✔stringr 1.3.0
## ✔readr   1.1.1     ✔forcats 0.3.0

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

## ── Conflicts ─────────────────────────────────────────────────────────────
──────────── tidyverse_conflicts() ──
## ✖dplyr::filter() masks stats::filter()
## ✖dplyr::lag()    masks stats::lag()
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.4.4

##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

```r
library(rgdal)
```

```
## Warning: package 'rgdal' was built under R version 3.4.4

## Loading required package: sp

## Warning: package 'sp' was built under R version 3.4.4

## rgdal: version: 1.3-4, (SVN revision 766)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.1.3, released 2017/20/01
```

```
##  Path to GDAL shared files:
/Users/darshanpatel/Library/R/3.4/library/rgdal/gdal
##  GDAL binary built with GEOS: FALSE
##  Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
##  Path to PROJ.4 shared files:
/Users/darshanpatel/Library/R/3.4/library/rgdal/proj
##  Linking to sp version: 1.3-1

library(RColorBrewer)
```

Question 1:

What is a census tract? How many census tracts are in New York County? (Provide the citations for references used.)

Answer: A census tract is an area of land considered to be a neighborhood by the Bureau of Census for analyzing populations (Source: www.census.gov). There are 288 census tracts in the New York County (Source: https://www.census.gov/geo/maps-data/).

Question 2:

Describe one advantage and one disadvantage of computing estimates after combining 5-years of data.

Answer: One advantage of computing estimates after combining 5-years of data is that it allows one to get a general sense of the estimate over a period of time. A disadvantage of this is that it can be affected by an outlier such as a piece of data that was high for one year only and remained roughly the same the other years.

Question 3:

Download the ACS data for 5-year estimates spanning from 2011 – 2015 for all New York County census tracts for the following variabes using American FactFinder. Table DP03 contains selected economic characteristics and Table DP04 includes selected housing characteristics. Each row in the table represents a single census tract in New York County.

- unemployment: Table DP03, variable HC03_VC07
- housing tenure (whether house is rented or owned): Table DP04, variable HC03_VC66
- no vehicles: Table DP04, variable HC03_VC85
- low occupancy: Table DP04, variable HC03_VC113 (transform this variable to get % overcrowded to use in the index)

Clean the data and merge the tables into one data frame, each row representing a census tract, each column representing one of the Townsend variables (keep the geography columns).

Answer:

```
# Import tables in its original form, making note of the NA values
dp03 <- read_delim("ACS_15_5YR_DP03_with_ann.csv", col_names = TRUE, delim =
",", na = "-")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character()
## )

## See spec(...) for full column specifications.

dp04 <- read_delim("ACS_15_5YR_DP04_with_ann.csv", col_names = TRUE, delim =
",", na = "-")

## Parsed with column specification:
## cols(
##   .default = col_character()
## )
## See spec(...) for full column specifications.

# Get only the rows where a census tract for NY county exists
dp03 <- filter(dp03, nchar(GEO.id) == 20)

## Warning: package 'bindrcpp' was built under R version 3.4.4

# Retain the 11 digit geography ids for map use in a later question
geo_ids <- dp03$GEO.id2

# Separate out the data in the 2 tables that is necessary for making the
Townsend index, including the geo label for labeling
dp03 <- select(dp03, 'GEO.display-label', HC03_VC07)
dp04 <- select(dp04, 'GEO.display-label', HC03_VC66, HC03_VC85, HC03_VC113)

# Join the datas together
x <- left_join(dp03, dp04)

## Joining, by = "GEO.display-label"

# Give columns reasonable names
x <- rename(x, geography_id = 'GEO.display-label',  unemployment = HC03_VC07,
            housing_tenure = HC03_VC66, no_vehicles = HC03_VC85,
            low_occupancy = HC03_VC113)

# Convert character columns into numeric columns for exploratory data
analysis and manipulation
x$unemployment <- as.numeric(x$unemployment)
x$housing_tenure <- as.numeric(x$housing_tenure)
x$no_vehicles <- as.numeric(x$no_vehicles)
x$low_occupancy <- as.numeric(x$low_occupancy)
x$low_occupancy <- 100 - x$low_occupancy
```

For each variable, construct a histogram and describe the shape of each histogram.

Answer:

```
# Create histogram for each variable
h_u <- ggplot(x, aes(unemployment)) + geom_histogram(color="black",
```

```
fill="coral1", na.rm = TRUE, binwidth = 1) +
  ggtitle("Unemployment Rate") + labs(x="Unemployment Rate", y = "Frequency")

h_ht <- ggplot(x, aes(housing_tenure)) + geom_histogram(color="black",
fill="deepskyblue2", na.rm = TRUE, binwidth = 2.5) +
  ggtitle("Housing Tenure") + labs(x="Housing Tenure", y = "Frequency")

h_nv <- ggplot(x, aes(no_vehicles)) + geom_histogram(color="black",
fill="darkolivegreen2", na.rm = TRUE, binwidth = 2.5) +
  ggtitle("People Having \n No Vehicles") + labs(x="No Vehicles", y =
"Frequency")

h_lo <- ggplot(x, aes(low_occupancy)) + geom_histogram(color="black",
fill="darkcyan", na.rm = TRUE, binwidth = 1) +
  ggtitle("Low Occupancy") + labs(x="Low Occupancy", y = "Frequency")

# Display the histogram in 1 grid
grid.arrange(h_u, h_ht, h_nv, h_lo)
```

The shape of the unemployment rate and low occupancy histograms is skewed right while the shape of the housing tenure and no vehicles is skewed left.

Compute the following summary statistics: mean, median, standard deviation, maximum and minimum for each variable.

Answer:

```
# Get the summary statistics of each column
ss_unemp <- x %>% summarize(mean(unemployment, na.rm = TRUE),
          median(unemployment, na.rm = TRUE), sd(unemployment, na.rm =
TRUE),
          max(unemployment, na.rm = TRUE), min(unemployment, na.rm = TRUE))
ss_ht <- x %>% summarize(mean(housing_tenure, na.rm = TRUE),
        median(housing_tenure, na.rm = TRUE), sd(housing_tenure, na.rm =
TRUE),
        max(housing_tenure, na.rm = TRUE), min(housing_tenure, na.rm = TRUE))
ss_nv <- x %>% summarize(mean(no_vehicles, na.rm = TRUE),
        median(no_vehicles, na.rm = TRUE), sd(no_vehicles, na.rm = TRUE),
        max(no_vehicles, na.rm = TRUE), min(no_vehicles, na.rm = TRUE))
ss_lo <- x %>% summarize(mean(low_occupancy, na.rm = TRUE),
        median(low_occupancy, na.rm = TRUE), sd(low_occupancy, na.rm = TRUE),
        max(low_occupancy, na.rm = TRUE), min(low_occupancy, na.rm = TRUE))
```

The summary statistics for unemployment are:

```
transpose(ss_unemp)
```

```
## [[1]]
## [[1]]$`mean(unemployment, na.rm = TRUE)`
## [1] 4.926855
##
## [[1]]$`median(unemployment, na.rm = TRUE)`
## [1] 4
##
## [[1]]$`sd(unemployment, na.rm = TRUE)`
## [1] 3.250082
##
## [[1]]$`max(unemployment, na.rm = TRUE)`
## [1] 25
##
## [[1]]$`min(unemployment, na.rm = TRUE)`
## [1] 0
```

The summary statistics for housing tenure are:

```
transpose(ss_ht)
```

```
## [[1]]
## [[1]]$`mean(housing_tenure, na.rm = TRUE)`
## [1] 77.53143
##
## [[1]]$`median(housing_tenure, na.rm = TRUE)`
## [1] 80.75
##
## [[1]]$`sd(housing_tenure, na.rm = TRUE)`
## [1] 19.1732
##
## [[1]]$`max(housing_tenure, na.rm = TRUE)`
```

```
## [1] 100
##
## [[1]]$`min(housing_tenure, na.rm = TRUE)`
## [1] 0
```

The summary statistics for no vehicles are:

```
transpose(ss_nv)
```

```
## [[1]]
## [[1]]$`mean(no_vehicles, na.rm = TRUE)`
## [1] 77.41357
##
## [[1]]$`median(no_vehicles, na.rm = TRUE)`
## [1] 78.6
##
## [[1]]$`sd(no_vehicles, na.rm = TRUE)`
## [1] 9.783584
##
## [[1]]$`max(no_vehicles, na.rm = TRUE)`
## [1] 100
##
## [[1]]$`min(no_vehicles, na.rm = TRUE)`
## [1] 0
```

The summary statistics for low occupancy are:

```
transpose(ss_lo)
```

```
## [[1]]
## [[1]]$`mean(low_occupancy, na.rm = TRUE)`
## [1] 5.883214
##
## [[1]]$`median(low_occupancy, na.rm = TRUE)`
## [1] 4.1
##
## [[1]]$`sd(low_occupancy, na.rm = TRUE)`
## [1] 5.164891
##
## [[1]]$`max(low_occupancy, na.rm = TRUE)`
## [1] 36.4
##
## [[1]]$`min(low_occupancy, na.rm = TRUE)`
## [1] 0
```

Question 4:

How many observations are missing for each variable?

Answer: For the unemployment feature, there are

```
# Count how many NAs there are in the unemployment column
as.numeric(count(filter(x, is.na(unemployment))))
```

```
## [1] 5
```

missing observations. For the housing tenure feature, there are

```
# Count how many NAs there are in the housing tenure column
as.numeric(count(filter(x, is.na(housing_tenure))))
```

```
## [1] 8
```

missing observations. For the no vehicles feature, there are

```
# Count how many NAs there are in the no vehicle column
as.numeric(count(filter(x, is.na(no_vehicles))))
```

```
## [1] 8
```

missing observations. For the low occupancy feature, there are

```
# Count how many NAs there are in the low occupancy column
as.numeric(count(filter(x, is.na(low_occupancy))))
```

```
## [1] 8
```

missing observations.

What percentage of census tracts do not have complete data? Is this a problem for our analysis? Justify your answer. (Note: do not delete tracts with missing data.)

Answer: The percentage of census tracts with incomplete data is

```
# Calculate how many distinct census tracts have a NA feature and the
percentage of it relative to the size of the total data
a <- union(which(is.na(x$unemployment)), which(is.na(x$housing_tenure)))
b <- union(which(is.na(x$no_vehicles)), which(is.na(x$low_occupancy)))
paste(round(100 * length(union(a,b)) / nrow(x), 3), "%", sep=' ')
```

```
## [1] "2.778 %"
```

Because this is a very small subset of the dataset, it will not be a problem for our analysis.

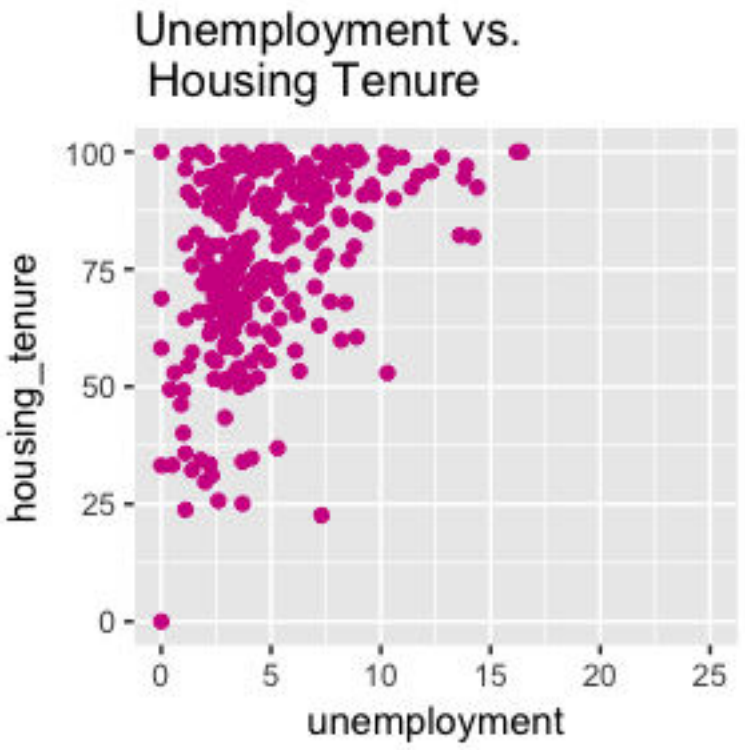
Question 5:

Construct scatterplots of the four variables.

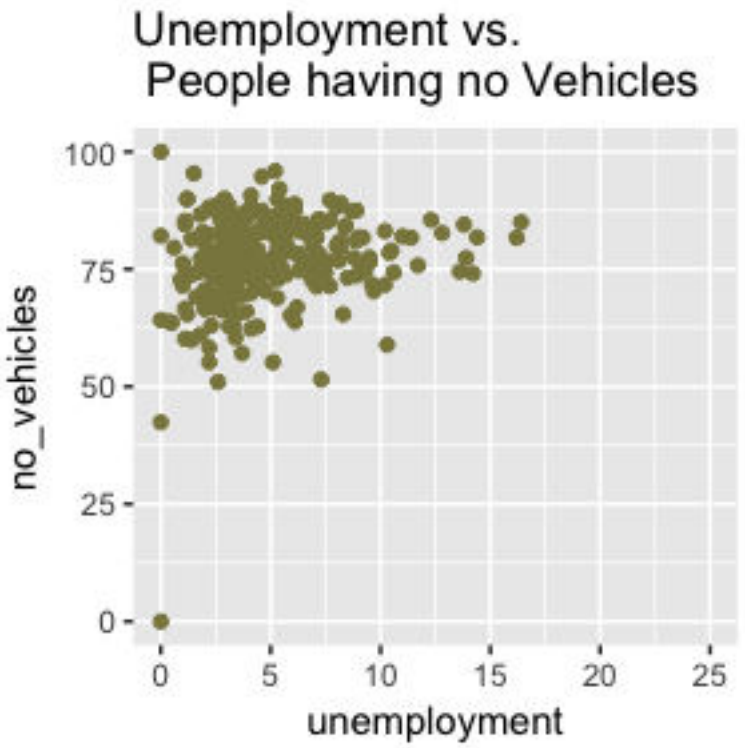
Answer:

```
# Scatterplots of each possible combinations
sp_u_ht <- ggplot(x, aes(unemployment, housing_tenure)) + geom_point(color =
"violetred", na.rm = TRUE) + ggtitle("Unemployment vs.\n Housing Tenure")

sp_u_nv <- ggplot(x, aes(unemployment, no_vehicles)) + geom_point(color =
"khaki4", na.rm = TRUE) + ggtitle("Unemployment vs. \n People having no
Vehicles")
```

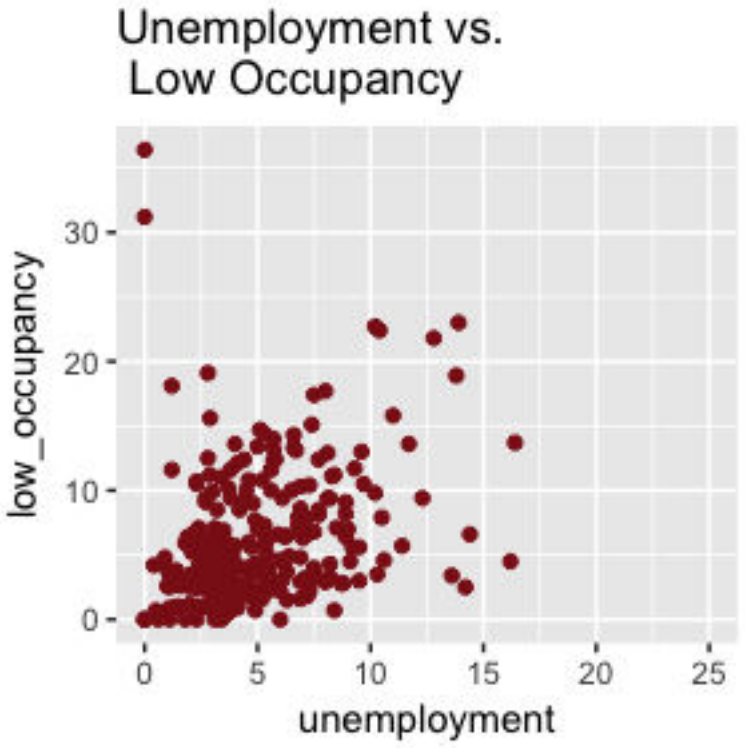
```
sp_u_lo <- ggplot(x, aes(unemployment, low_occupancy)) + geom_point(color =
"firebrick4", na.rm = TRUE) + ggtitle("Unemployment vs. \n Low Occupancy")

sp_ht_nv <- ggplot(x, aes(housing_tenure, no_vehicles)) + geom_point(color =
"goldenrod4", na.rm = TRUE) + ggtitle("Housing Tenure vs. \n People having no
Vehicles")

sp_ht_lo <- ggplot(x, aes(housing_tenure, low_occupancy)) + geom_point(color
= "blue2", na.rm = TRUE) + ggtitle("Housing Tenure vs. \n Low Occupancy")

sp_nv_lo <- ggplot(x, aes(no_vehicles, low_occupancy)) + geom_point(color =
"chocolate", na.rm = TRUE) + ggtitle("People Having No Vehicles vs. \n Low
Occupancy")

# Show plots
sp_u_ht
```



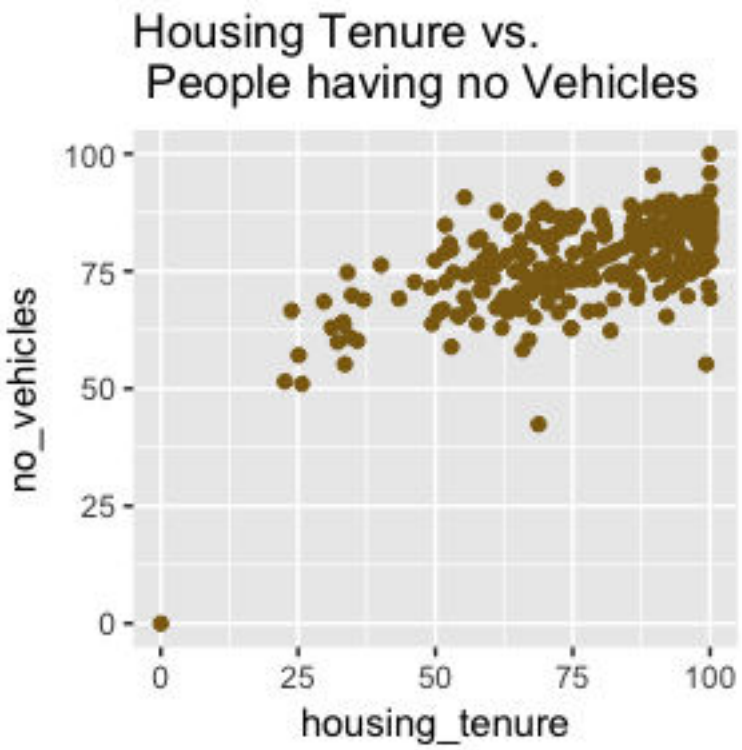

```
sp_u_nv
```



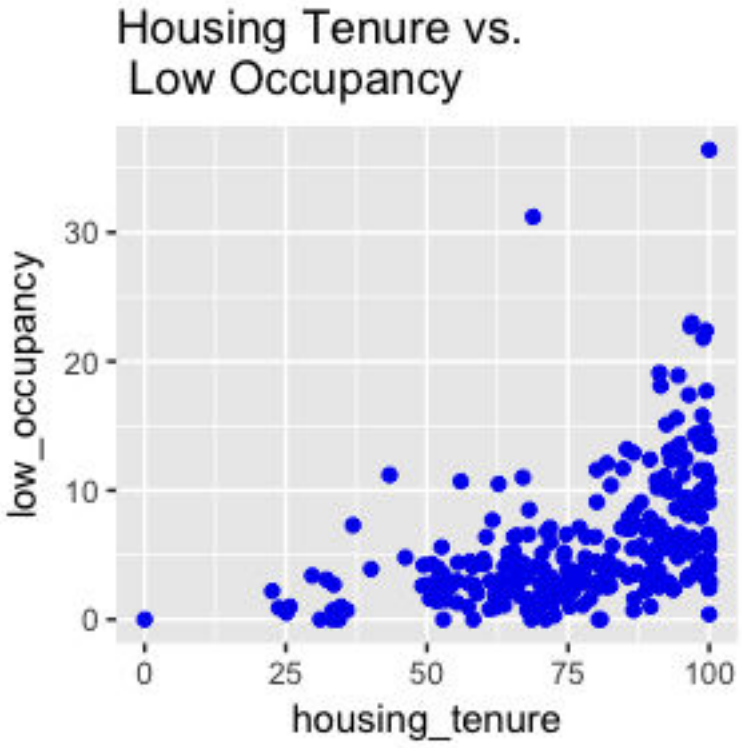

```
sp_u_lo
```



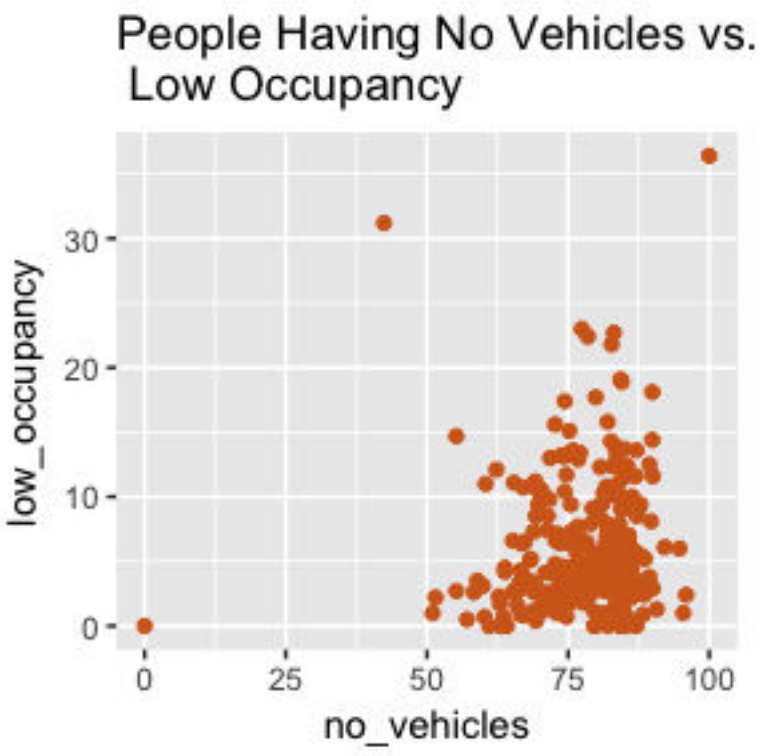

```
sp_ht_nv
```




```
sp_ht_lo
```




```
sp_nv_lo
```



Are they linearly related?

Answer: According to the scatterplots,

- unemployment and housing tenure are linearly related
- unemployment and no vehicles are linearly related
- unemployment and low occupancy are linearly related
- housing tenure and no vehicles are linearly related
- housing tenure and low occupancy are linearly related
- no vehicles and low occupancy are not linearly related

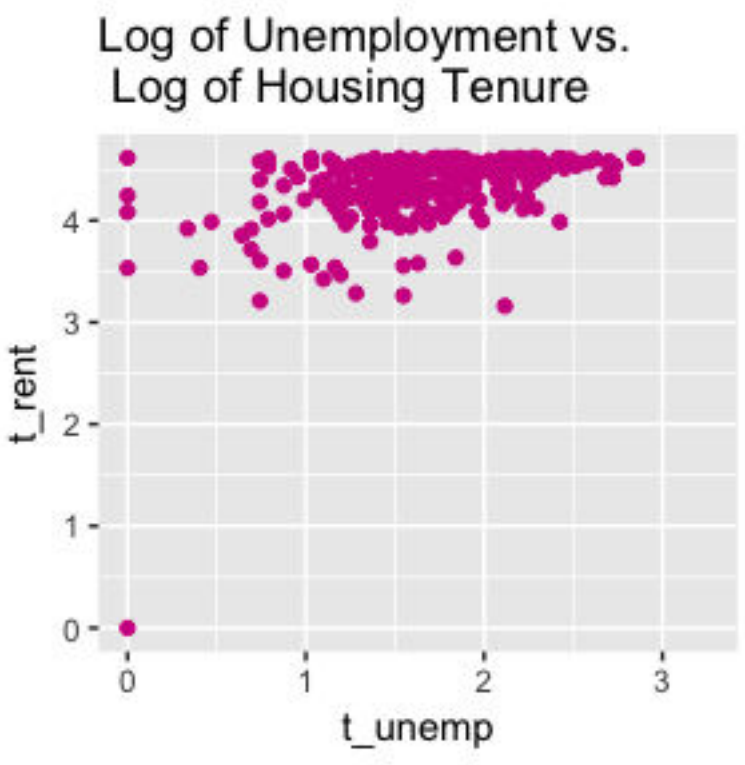
Now, transform the variables as given in step (a), adding the transformed variables to the data frame.

```
# Transformation of the variables
t_unemp <- log(x$unemployment + 1)
t_oc <- log(x$low_occupancy + 1)
t_rent <- log(x$housing_tenure + 1)
t_car <- sqrt(x$no_vehicles)

# Bind the transformed variables together into a tibble and bind this tibble
to the original data tibble
t_vars <- cbind(t_unemp, t_oc, t_rent, t_car)
x <- as.tibble(cbind(x, t_vars))
```

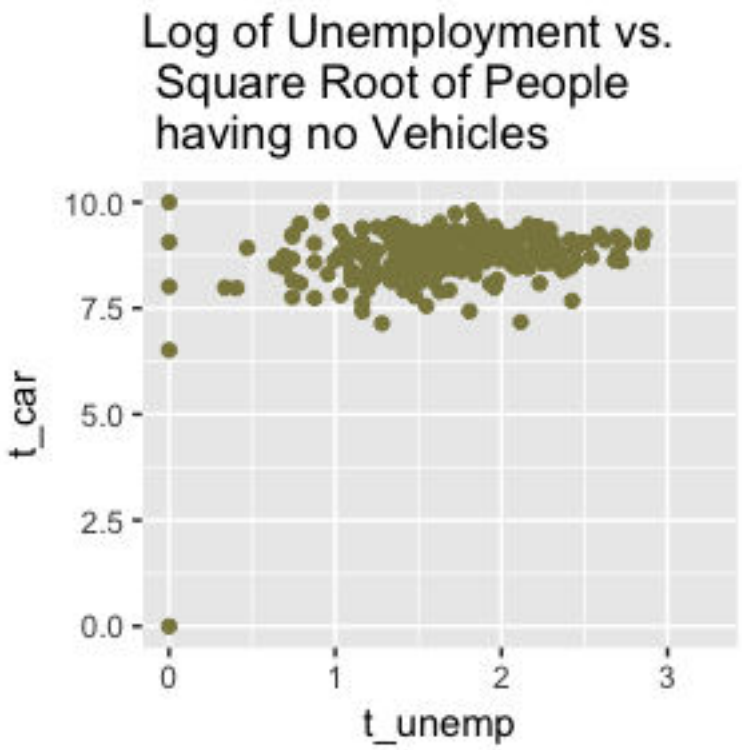
Make another scatter plot matrix with the transformed variables.

Answer:

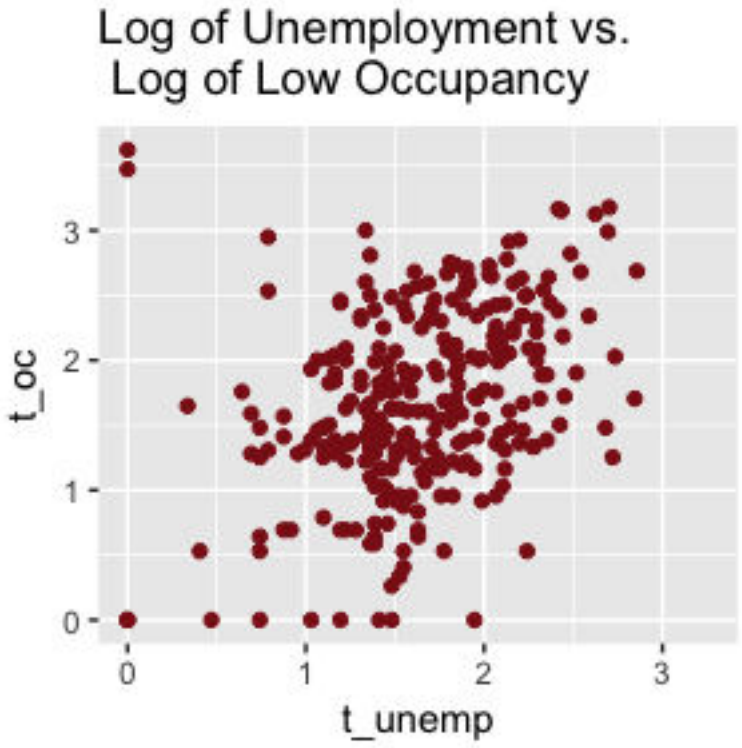
```r
# Scatterplots of every possible combination of transformed variables
tsp_u_r <- ggplot(x, aes(t_unemp, t_rent)) + geom_point(color = "violetred",
na.rm = TRUE) + ggtitle("Log of Unemployment vs. \n Log of Housing Tenure")

tsp_u_c <- ggplot(x, aes(t_unemp, t_car)) + geom_point(color = "khaki4",
na.rm = TRUE) + ggtitle("Log of Unemployment vs. \n Square Root of People \n
having no Vehicles")

tsp_u_oc <- ggplot(x, aes(t_unemp, t_oc)) + geom_point(color = "firebrick4",
na.rm = TRUE) + ggtitle("Log of Unemployment vs. \n Log of Low Occupancy")

tsp_r_c <- ggplot(x, aes(t_rent, t_car)) + geom_point(color = "goldenrod4",
na.rm = TRUE) + ggtitle("Log of Housing Tenure vs. \n Square Root of People
\n having no Vehicles")

tsp_r_oc <- ggplot(x, aes(t_rent, t_oc)) + geom_point(color = "blue2", na.rm
= TRUE) + ggtitle("Log of Housing Tenure vs. \n Log of Low Occupancy")

tsp_c_oc <- ggplot(x, aes(t_car, t_oc)) + geom_point(color = "chocolate",
na.rm = TRUE) + ggtitle("Square Root of People \n Having No Vehicles vs. \n
Log of Low Occupancy")

# Show plots
tsp_u_r
```



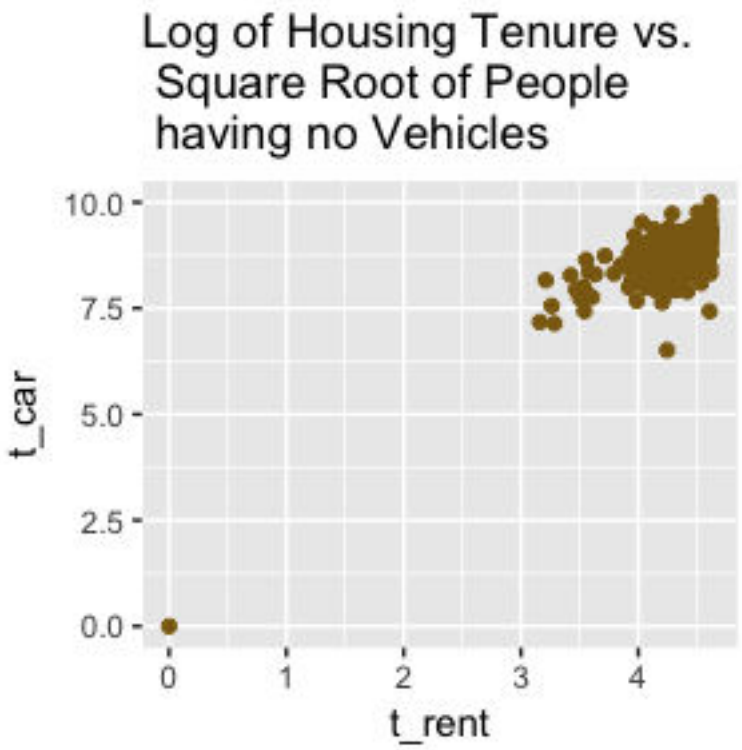

```r
tsp_u_c
```



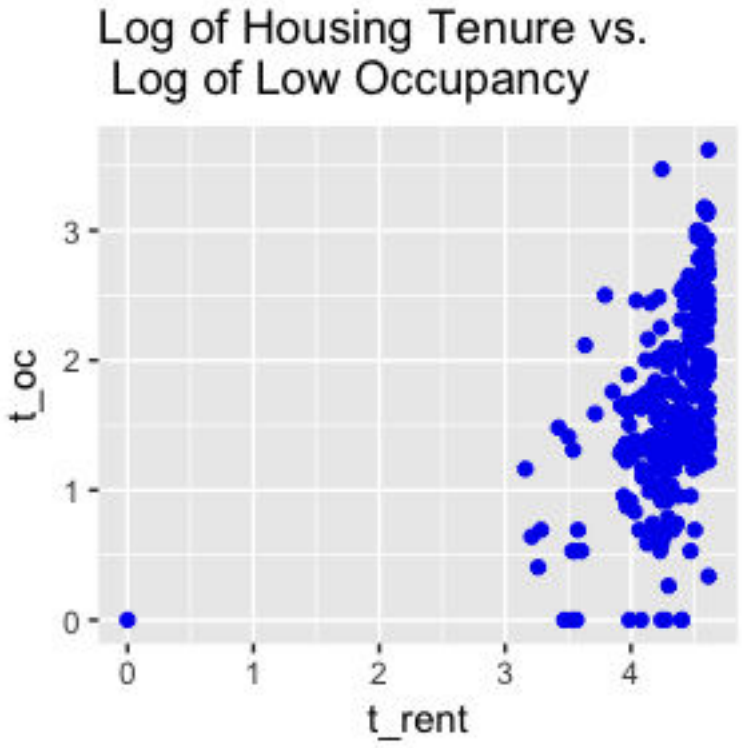

```
tsp_u_oc
```



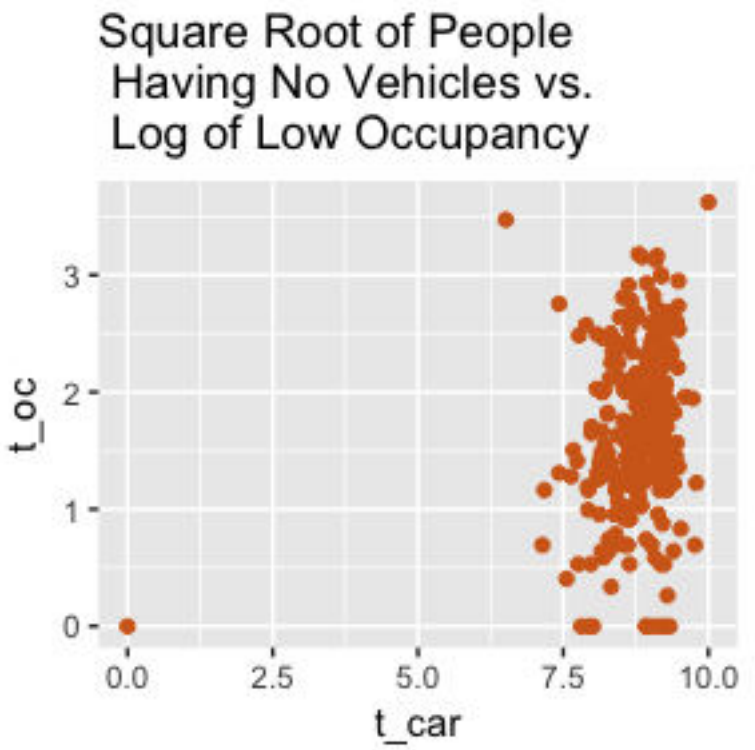

```
tsp_r_c
```




```
tsp_r_oc
```




```
tsp_c_oc
```



Are they linearly related?

Answer: According to the scatterplots,

- log of unemployment and log of housing tenure are linearly related
- log of unemployment and square root of no vehicles are linearly related
- log of unemployment and log of low occupancy are not linearly related
- log of housing tenure and square root of no vehicles are not linearly related
- log of housing tenure and log of low occupancy are linearly related
- square root of no vehicles and log of low occupancy are linearly related

Construct a correlation matrix of the transformed variables and describe your results.

Answer:

```
cor(t_vars, use = "complete.obs")
```

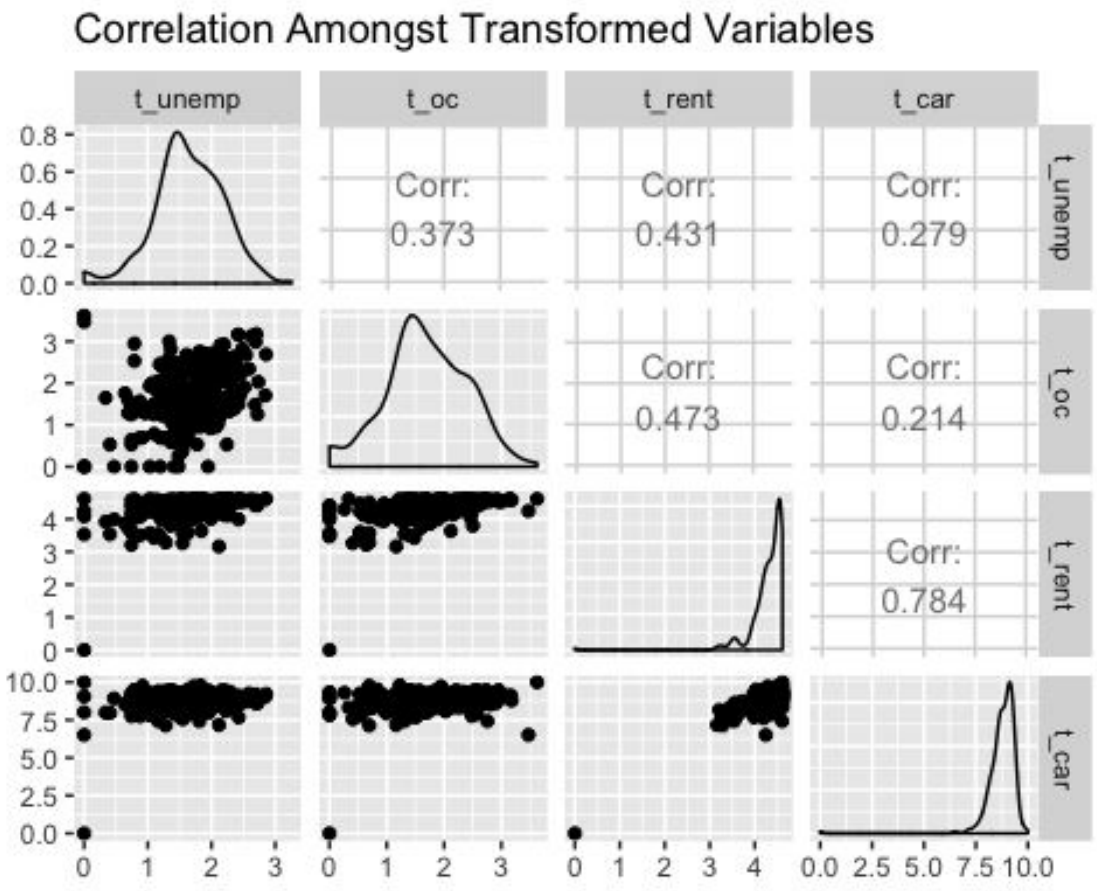
```
##           t_unemp      t_oc    t_rent     t_car
## t_unemp 1.0000000 0.3725944 0.4311138 0.2788261
## t_oc    0.3725944 1.0000000 0.4730620 0.2141658
## t_rent  0.4311138 0.4730620 1.0000000 0.7842310
## t_car   0.2788261 0.2141658 0.7842310 1.0000000
```

There is a high correlation of 0.78 between t_{car} and t_{rent}. t_{rent} also shares a moderately lower correlation with t_{unemp} and t_{oc}, of 0.43 and 0.47 respectively. t_{car} shares a low correlation with t_{unemp} and t_{oc}, of 0.27 and 0.21 respectively. All of this indicates that square root of people having cars and the log of housing tenure have more correlation with each other compared to the other combinations of other variables.

The above results can also be seen with one line of code and interpreted similarly.

```
ggpairs(data.frame(t_vars), progress = FALSE, title = "Correlation Amongst
Transformed Variables")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 8 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 8 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 8 rows containing missing values

## Warning: Removed 8 rows containing missing values (geom_point).

## Warning: Removed 8 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 8 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 8 rows containing missing values

## Warning: Removed 8 rows containing missing values (geom_point).

## Warning: Removed 8 rows containing missing values (geom_point).

## Warning: Removed 8 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 8 rows containing missing values

## Warning: Removed 8 rows containing missing values (geom_point).

## Warning: Removed 8 rows containing missing values (geom_point).

## Warning: Removed 8 rows containing missing values (geom_point).

## Warning: Removed 8 rows containing non-finite values (stat_density).
```



Question 6:

Compute the Townsend index value for each census tract. Add the index to your data frame.

Answer:

```
# Standardize each variable and bind them together to a new matrix
std_vars <- cbind((t_unemp - mean(t_unemp, na.rm = TRUE)) / sd(t_unemp, na.rm
= TRUE), (t_oc - mean(t_oc, na.rm = TRUE)) / sd(t_oc, na.rm = TRUE),
      (t_rent - mean(t_rent, na.rm = TRUE)) / sd(t_rent, na.rm = TRUE),
(t_car - mean(t_car, na.rm = TRUE)) / sd(t_car, na.rm = TRUE))

# Add the sum of the standardized variables, or the Townsend Index, to the
data tibble
x <- as.tibble(cbind(x, rowSums(std_vars, na.rm = FALSE)))

# Name the new column Townsend Index
x <- rename(x, Townsend_Index = "rowSums(std_vars, na.rm = FALSE)")
```

For how many census tracts are you able to compute the Townsend index? Why does this number not equal the total number of census tracts?

Answer:

```r
# Get number of non-null values in the Townsend Index column
length(which(!is.na(x$Townsend_Index)))
```

```
## [1] 280
```

There are 280 census tracts where the Townsend index was able to be calculated. This does not equal the total number of census tracts because of the missing data. If a census tract had a missing data in one or more of its variables, then its Townsend index is not computable.

Question 7:

Identify which census tract is the most deprived and which is the least deprived (give the census tract number and deprivation index level). Based on the results, would you like to live in the least deprived census tract? Justify your answer.

Answer:

```r
# Create table of census tract that are most and least deprived showing only
its tract number and Townsend Index
x[c(which.min(x$Townsend_Index),which.max(x$Townsend_Index)),
c("geography_id", "Townsend_Index")]
```

```
## # A tibble: 2 x 2
##   geography_id                                    Townsend_Index
##   <chr>                                                    <dbl>
## 1 Census Tract 217.03, New York County, New York          -28.6
## 2 Census Tract 285, New York County, New York               5.04
```

The census tract that is least deprived is Census Tract 217.03 and its deprivation level is −28.58582. The census tract that is most deprived is Census Tract 285 and its deprivation index level is 5.044114. Based on the results, I would live in the least deprived census tract because its Townsend Index is several values away from the most deprived tract's Townsend Index.

Question 8:

The ACS data includes not only estimates but their margins of errors which was ignored in our calculations. What are the implications?

Answer: The implications of ignoring the margins of errors in our calculations is that the Townsend index calculated for each census tract may not be an accurate value for that specific census tract. Rather, it may lie in a range of the margin of error. This is because the margin of error is interpreted as the standard error in a 90% confidence interval from the sampled variable, according to the notes file from the dataset. This could mean that Census Tract A that is more deprived than Census Tract B, with similar Townsend index, could actually be less deprived than Census Tract B or vice versa.

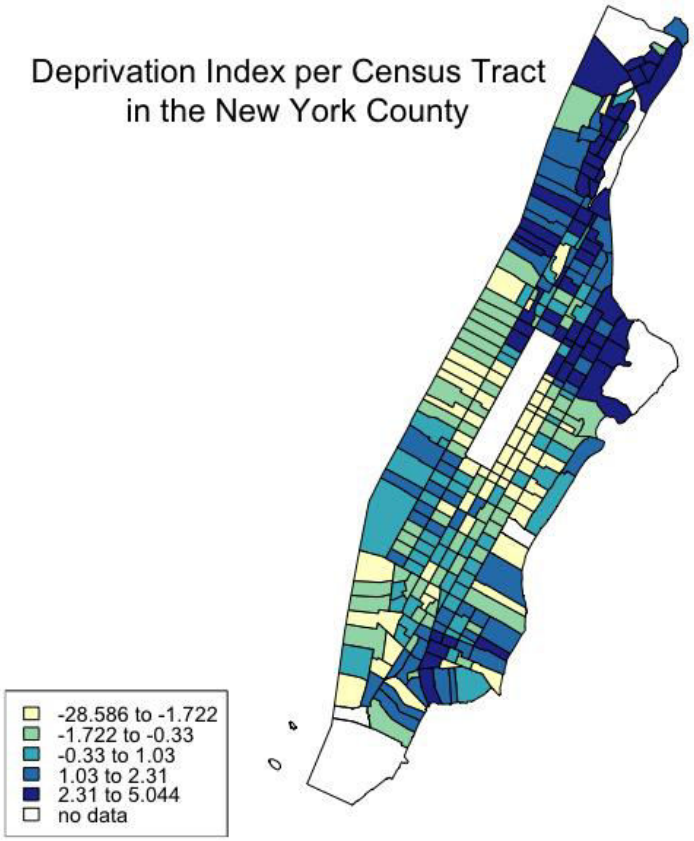
Question 9:

Construct a map color-coded by the deprivation index value quintile. Each quintile (i.e., 20$) should be assigned a different color from least to most deprived. Download the shape

files for New York state census tracts for 2015 from the US Census Bureau website. Extract the tracts for New York County only. Include a legend and plot title.

Answer:

```
# Import the spatial file
county_map <- readOGR(dsn = "tl_2015_36_tract", layer = "tl_2015_36_tract")

## OGR data source with driver: ESRI Shapefile
## Source: "/Users/darshanpatel/Library/Mobile
Documents/com~apple~CloudDocs/Shared Files/SD 7844/HWs/HW
2/tl_2015_36_tract", layer: "tl_2015_36_tract"
## with 4918 features
## It has 12 fields
## Integer64 fields read as strings:  ALAND AWATER

# Filter out the non-New York County area
county_map <- subset(county_map, is.element(county_map$GEOID, geo_ids))

# Join the geo ids with its respective Townsend Index
mapping <- data.frame(geo_ids, x$Townsend_Index)

# Get the quintiles for the Townsend Index values
quintiles <- round(quantile(mapping$x.Townsend_Index, probs = seq(0,1,0.2),
na.rm = TRUE), digits = 3)

# Create a vector to be used in a map legend using the quintiles calculated
key <- c()

for(i in 1:5){
    key <- c(key, paste(quintiles[i], quintiles[i+1], sep = " to "))
}

# Determine which Townsend Index and its geo id belong in which quintile
region
mapping <- mapping %>% mutate(quintile = cut(mapping$x.Townsend_Index, breaks
= quintiles, include.lowest = TRUE, labels = brewer.pal(n = 5, "YlGnBu")))

# Match the geo ids in the spatial file with the ones in the data set
mapping <- mapping %>% slice(match(county_map$GEOID, mapping$geo_ids))

# Plot the map with its appropriate Townsend Index quintile and add a legend
and title
plot(county_map, col = as.character(mapping$quintile))
legend("bottomleft", legend = c(key, "no data"), fill = c(brewer.pal(n = 5,
"YlGnBu"), "white"), cex=0.9, y.intersp = 0.8)
text(-74.03921, 40.86083, cex = 1.4, labels = c("Deprivation Index per Census
Tract \n in the New York County"))
```



Describe the patterns you see, especially in relation to what you know about neighborhoods in New York City. What does the large rectangle in the middle of the map represent?

Answer: A positive deprivation index indicates an area is more deprived than one that is negative or zero. According to this map, most deprived areas are situated in Harlem, Upper Manhattan and the Lower East Side. The less deprived areas are situated in Upper West Side, Upper East Side, Murray Hill and Greenwich Village. The areas that are at the average level of deprivation consist of Chelsea, Hell's Kitchen and parts of Chinatown. The large rectangle in the middle of the map represents Central Park.

Question 10:

In which census tract is 140 W. 62nd St.? What is the deprivation level rank (where a rank of 1 is the most deprived)? (Provide citations for references.)

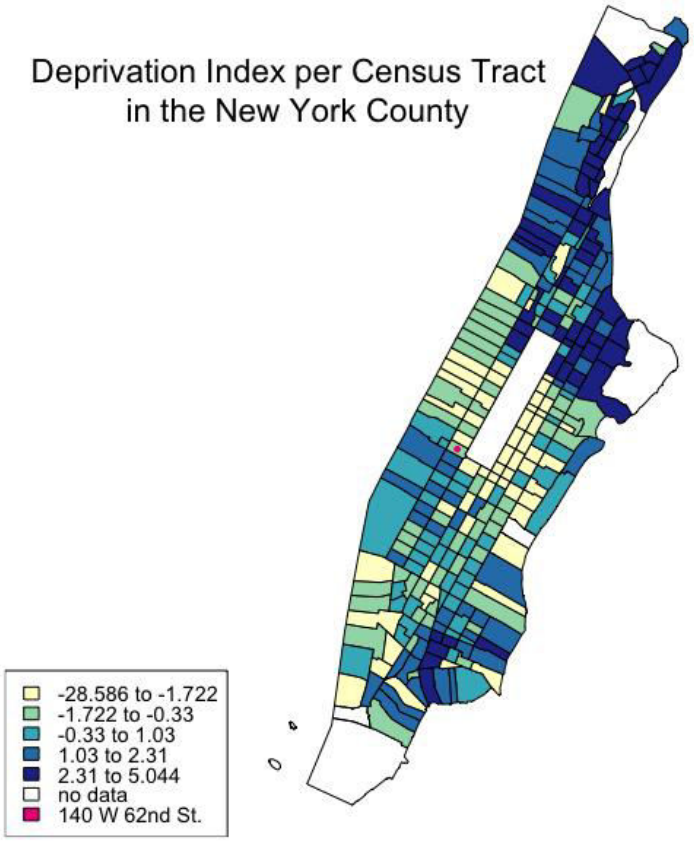
Answer: 140 W. 62nd St. is located in Census Tract 145, according to the Geocoder on the US Census website (https://geocoding.geo.census.gov/geocoder/). Its deprivation rank is

```
# Get the row index where Census Tract 145 is stored
tract <- grep(145, x$geography_id)

# Get the location of Census Tract 145's Townsend Index in a sorted list and
compute its rank
match(x[tract,]$Townsend_Index, sort(x$Townsend_Index)) / nrow(x)

## [1] 0.21875
```

Mark it on the map and add it to the legend.

Answer:

```
# Plot the map with legend and title
plot(county_map, col = as.character(mapping$quintile))
legend("bottomleft", legend = c(key, "no data", "140 W 62nd St."), fill =
c(brewer.pal(n = 5, "YlGnBu"), "white", "deeppink2"), cex=0.9, y.intersp =
0.8)
text(-74.03921, 40.86083, cex = 1.4, labels = c("Deprivation Index per Census
Tract \n in the New York County"))

# Add the point where 140 W. 62nd St. is located
points(coordinates(subset(county_map, county_map$GEOID == 36061014500)),
cex=0.5, pch=19, col = "deeppink2")
```



Question 11:

New York County is an urban county, however New York state has roughly 22 counties classified as rural (e.g., Allegany, Essex, Otsego, Sullivan). Would it make sense to compute the Townsend index values for all census tracts within New York state combined? Why or why not?

Answer: It would not make sense to compute the Townsend index values for all census tracts within the New York state combined. New York County is an urban county; it may be less deprived than the rest of the state, including the 22 rural counties. It may be that New York county is an outlier when looking at Townsend index values since it is not a rural area like the other 22 counties.