Assignment 2

Himani Anil Deshpande

Load tidyCensus

```
library(tidycensus)
library(tidyverse)

census_api_key("2103368d35e640a56f8d434964835a51bea42115")
```

```
## To install your API key for use in future sessions, run this function with `install = TRUE`.
```

Load Vehicle Data and variables

```
vehicles = get_acs(geography = "county",
  variables = c(total = "B08201_001",
median_income = "B19013_001",
cars0 = "B08201_002",
cars1 = "B08201_003",
cars2 = "B08201_004",
cars3 = "B08201_005",
cars4 = "B08201_006"),
year = 2019)
```

```
## Getting data from the 2015-2019 5-year ACS
```

Create a new Column "NEwEstimate" which calculates the total number of vehicles in each household by multiplying the number of households with the number of cars it has.

```
vehicle_new_estimates = transform(vehicles,
        NewEstimate=ifelse(variable=='cars0', estimate*0,
         ifelse(variable=='cars2', estimate*2,
        ifelse(variable=='cars3', estimate*3,
               ifelse(variable=='cars4' ,estimate*4, estimate)))))
```

Creating a new Column which will sum up the number of Cars in a county.(i.e sum up NewEstimate of variables cars0, cars1, cars2, cars3, cars4)

```
#aggregate(vehicle_new_estimates$NewEstimate,
#by=list(variable=vehicle_new_estimates$variable %in%
#c('cars0','cars1', 'cars2', 'cars3', 'cars4')), FUN=sum)

library(dplyr)
library(stringr)
#vehicle_new_estimates %>%
 # group_by(GEOID == "01001", ) %>%
  #summarise(Mean = sum(NewEstimate))

result <- vehicle_new_estimates %>%
group_by(NAME, variable %in% c('cars0','cars1', 'cars2', 'cars3', 'cars4'),
         variable == 'total', variable =='median_income') %>%
        mutate(TotalVehicles = sum(NewEstimate))

  #summarise(NewEstimate,across(everything(), sum))
   #summarise(Mean = sum(NewEstimate), groups(NAME,
#variable %in% c('cars0','cars1', 'cars2', 'cars3', 'cars4')))
#result
```

Create a wide data set of vehicles to work on

I have filtered the variable column to 3 types(cars0, total and median_income). The result tibble from above has same "total number of vehicles" in cars0, cars1, cars2, cars3 and cars4 due to grouping on these variables.

Then I have selected the columns which I will be working on (GEOID, Name(of county), variable, total vehicles) and spread the dataset into wide format on basis of variable(column Name) and Total Vehicles(value).

When we spread the data we get the columns as GEOID, Name(of county), cars0, total, median_income. Then for readability I have renamed cars0 and TotalVehicles (as explained, cars0 containts the total number of variables in a county)

```
wide_data_set = filter(result,variable %in%
  c("cars0", "total","median_income"))[,c("GEOID", "NAME", "variable", "TotalVehicles")] %>%
  spread(variable , TotalVehicles)  %>%
  rename(TotalVehicles = cars0)
#wide_data_set
```

Ploting distribution of response variable

By transforming the data to "wide form" or otherwise, estimate the mean number of vehicles owned per household for each county in the data set. Plot the distribution of this variable.

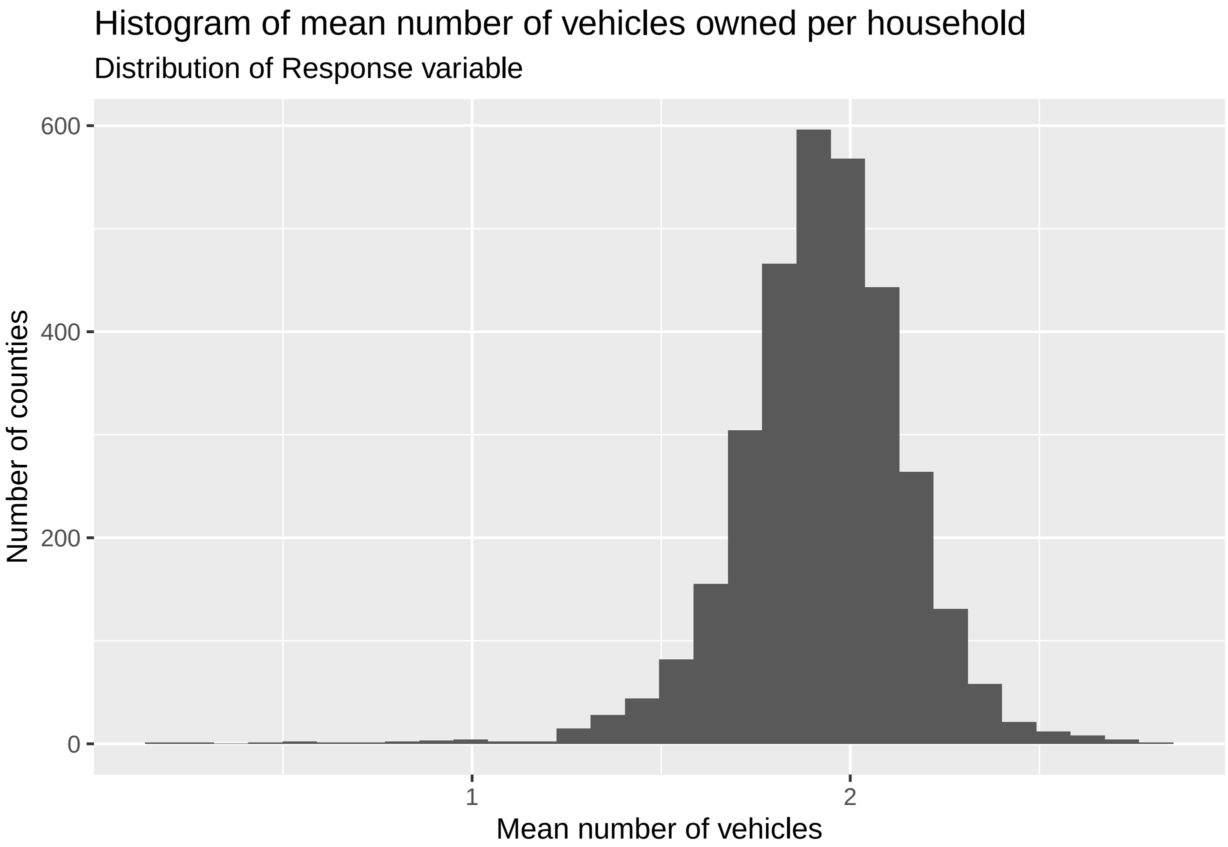
I have converted the original data into wide form as required in the previous section. The response variable is the mean number of vehicles owned per household for each county in the data set.

Question 1

```
ggplot(wide_data_set, aes(x= TotalVehicles/total))  +
  geom_histogram()+
  ggtitle("Histogram of mean number of vehicles owned per household")+
  labs(subtitle = "Distribution of Response variable")+

  xlab("Mean number of vehicles")+
  ylab("Number of counties")
```

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



The histogram shows average number of cars that counties contain. (Looks like it is negatively skewwed)

Question 2

I have plotted the vehicles owned (y-axis) against median income (x-axis) for each county. A regression line is a straight line that de- scribes how a response variable y changes as an explanatory variable x changes.

The regression line fits the data points but its not the best representation of relationship between y and x. This line has more number of outliers. However, it could be improved by performing transformations on data. The slope of the line could be increased by using transformations to better fit the data points which represent the counties which have mean number of vehicles in the range of [1,2]

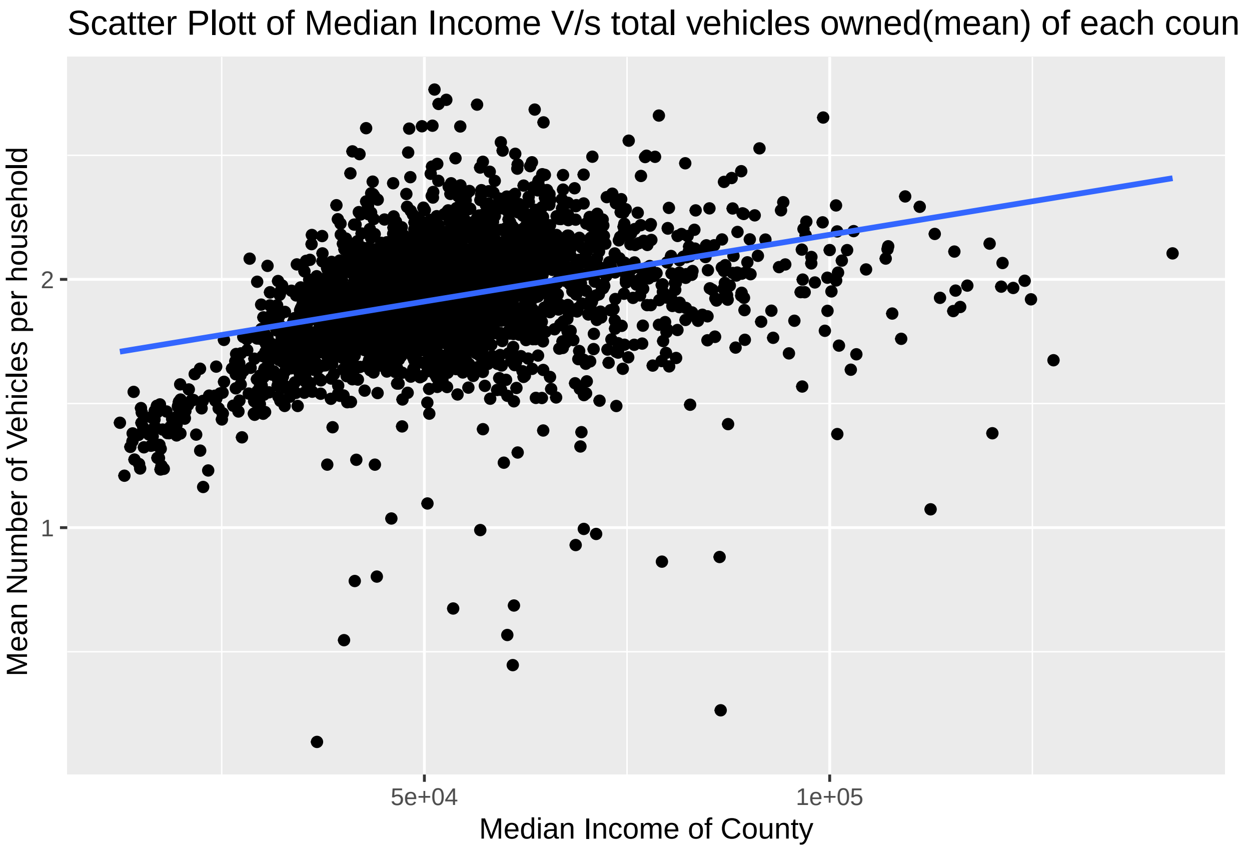
```
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

```
ggplot(wide_data_set, aes( x = median_income, y =  TotalVehicles/total)) +
  geom_jitter()  +
  geom_smooth(method = "lm", se= FALSE) +
  #geom_smooth(method = "rlm", se= FALSE, method.args = list(psi = psi.bisquare),color= "orange") +
  ggtitle("Scatter Plott of Median Income V/s total vehicles owned(mean) of each county")+
  ylab(" Mean Number of Vehicles per household")+
  xlab("Median Income of County")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



## Question 3

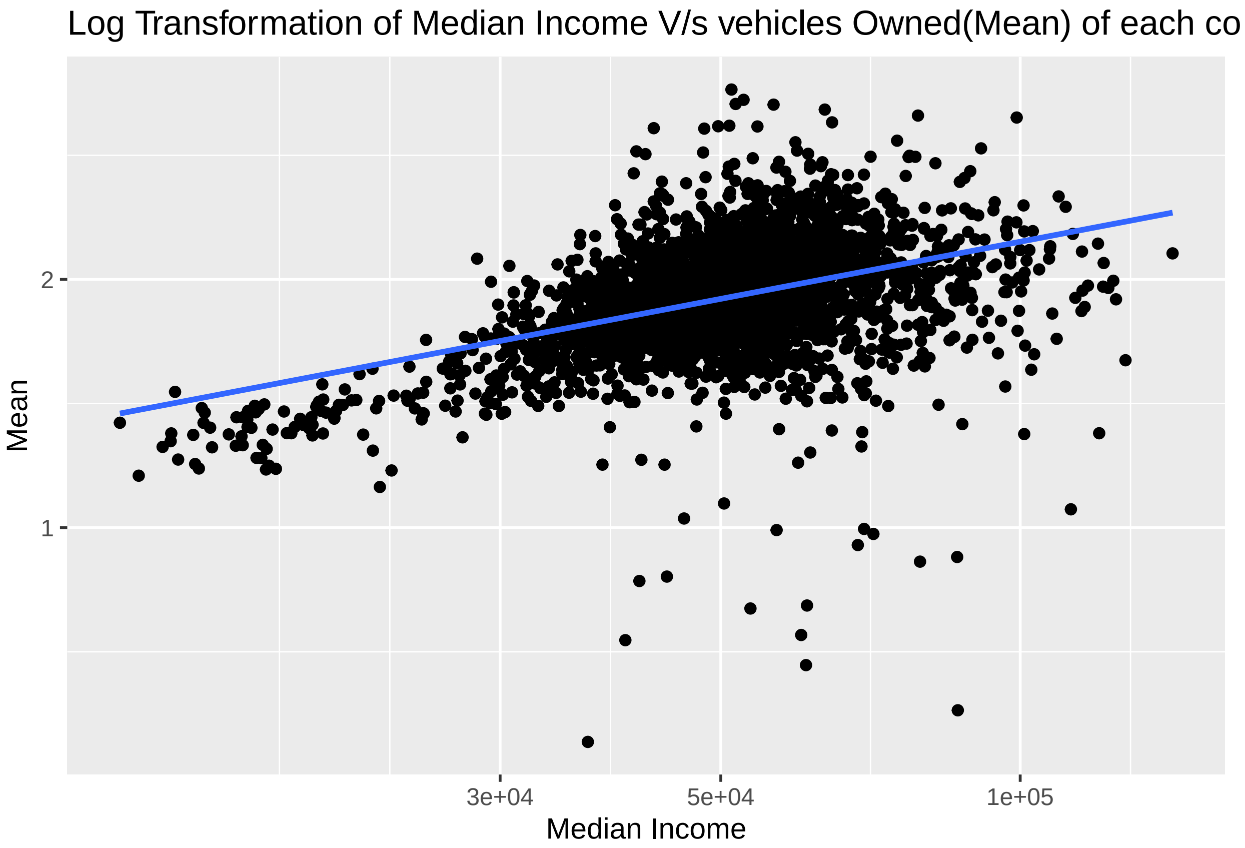
Log transformation on independent Variable

I have plotted vehicles owned against median income using a log scale on the x-axis for each county. After performing log transformation on the explanatory variable(median income of county). This has spread the dense cluster across the plot compared to the previous plot, where the cluster was on one side. The nnumber of outliers seems to have decreased but not significantly

```
ggplot(wide_data_set, aes( x = median_income, y =  TotalVehicles/total)) +
  geom_jitter()  +
  scale_x_log10() +
  geom_smooth(method = "lm", se= FALSE) +
  ggtitle("Log Transformation of Median Income V/s vehicles Owned(Mean) of each county") +

  ylab(" Mean ")+
  xlab("Median Income")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



## Question 4

I Have fitted the wide vehicle data in a linear regression model which will predict the vehicles owned using median income for each county

```
wide_data_set.lm = lm(TotalVehicles/total ~ median_income, data = wide_data_set)
#augment(wide_data_set.lm)
wide_data_set.lm.df = data.frame(wide_data_set, .resid  = residuals(wide_data_set.lm ),
                         .fitted = fitted(wide_data_set.lm))
#wide_data_set.lm.df
```

```
#wide_data_set.lm.df[,c("GEOID", "NAME", ".resid")] %>% arrange(.resid)
#ggplot(wide_data_set.lm.df, aes(x = median_income , y = .resid)) +
 # geom_jitter(width =0.5, size =0.7) +
  #geom_abline(slope = 0)
```

Listing the ten counties with the lowest (most negative) residuals

```
wide_data_set.lm.df[,c("GEOID", "NAME", ".resid")] %>% slice_min(.resid, n = 10)
```

```
##      GEOID                                    NAME    .resid
## 1859 36061                 New York County, New York -1.843270
## 82   02158              Kusilvak Census Area, Alaska -1.702686
## 87   02188          Northwest Arctic Borough, Alaska -1.523298
## 1852 36047                    Kings County, New York -1.398379
## 1831 36005                    Bronx County, New York -1.310750
## 85   02180                  Nome Census Area, Alaska -1.283944
## 71   02050                Bethel Census Area, Alaska -1.255656
## 320  11001 District of Columbia, District of Columbia -1.225095
## 86   02185                 North Slope Borough, Alaska -1.205484
## 224  06075              San Francisco County, California -1.173058
```

```
#ggplot(wide_data_set.lm.df, aes(x = median_income , y = .resid)) +
 # geom_jitter(width =0.5, size =0.7) +
  #geom_abline(slope = 0)
```

Negative residuals shows that we have over predicted. It means the actual value was less than the predicted value.

According to above list, if I scrutinize the counties that have been mentioned, we see that these counties do have very few number of cars. For example, the counties in Alaska barely have any cars. At most there are 2200 cars in their county with a median income around 50k to 70k. However, there are counties like NY and San Fransico which shows that they have more than 5M cars and median income in the range of 40K to 1.2M. There are a fair number of households which dont use cars, so it seems just on the basis of the given variables we can say that they are good places to live. However we havent taken into account other factors like public transport frequency, availability of stores nearby and area of the city,etc. Without these we cant make a solid conclusion of whether we need a car or not to live in these counties.

Question 5

I have fited a linear regression that predicts vehicles owned using log median income for each county.

```
wide_data_set.lm1 = lm(TotalVehicles/total ~ log( median_income), data = wide_data_set)
#augment(wide_data_set.lm)
wide_data_set.lm1.df = data.frame(wide_data_set, .resid  = residuals(wide_data_set.lm1 ),
                         .fitted = fitted(wide_data_set.lm1))
#wide_data_set.lm1.df
```

List the ten counties with the lowest (most negative) residuals.

```
wide_data_set.lm1.df[,c("GEOID", "NAME", ".resid")] %>% slice_min(.resid, n = 10)
```

```
##      GEOID                                       NAME    .resid
## 1859 36061                     New York County, New York -1.839540
## 82   02158                  Kusilvak Census Area, Alaska -1.682535
## 87   02188              Northwest Arctic Borough, Alaska -1.540889
## 1852 36047                        Kings County, New York -1.415902
## 85   02180                      Nome Census Area, Alaska -1.301545
## 1831 36005                        Bronx County, New York -1.301489
## 71   02050                    Bethel Census Area, Alaska -1.270095
## 320  11001 District of Columbia, District of Columbia -1.221571
## 86   02185                   North Slope Borough, Alaska -1.211717
## 224  06075              San Francisco County, California -1.116890
```

```
#ggplot(wide_data_set.lm.df, aes(x = median_income , y = .resid)) +
 # geom_jitter(width =0.5, size =0.7) +
  #geom_abline(slope = 0)
```

It seems that the result havent changed even after taking a log of the median income. So the same explanation as above applies.