California Income Tax Exploratory Data Analysis by Anjali Godbole

Introduction

A data set from Kaggle relating income data to US zipcodes was explored. This analysis focused on data from California. In addition, a subset of variables were selected for analysis. Please see the readme file for a link to the data set and documentation. All dollar amounts are in units of thousands of dollars.

Univariate Plots Section

The data types and summary statistics for the data set were first examined.

```
## 'data.frame':    8903 obs. of  20 variables:
##  $ State             : Factor w/ 1 level "CA": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Zipcode           : int  90001 90001 90001 90001 90001 90001 90002 90002 90002
90002 ...
##  $ AGI_category      : int  1 2 3 4 5 6 1 2 3 4 ...
##  $ Num_returns       : int  13020 6110 1630 420 170 20 12010 5330 1430 380 ...
##  $ Num_single_returns: int  7130 1810 370 70 30 0 6500 1660 350 80 ...
##  $ Num_joint_returns : int  1880 2010 810 260 120 0 1410 1530 620 210 ...
##  $ Num_dependents    : int  14420 9520 2660 660 250 30 13620 7940 2290 610 ...
##  $ AGI               : int  179555 211459 97566 35107 20614 7398 164428 185387 853
70 31894 ...
##  $ Num_ord_div       : int  60 60 50 0 20 0 50 50 40 30 ...
##  $ Amt_ord_div       : int  52 46 31 0 68 0 19 90 41 6 ...
##  $ Num_bus_inc       : int  3530 560 130 30 30 0 3690 540 140 60 ...
##  $ Amt_bus_inc       : int  31720 6746 1359 582 528 0 34580 5298 1597 753 ...
##  $ Num_cap_gain      : int  60 60 30 20 30 0 50 50 40 30 ...
##  $ Amt_cap_gain      : int  16 123 137 188 296 0 -11 70 3 70 ...
##  $ Num_tax_paid      : int  450 1190 740 260 130 20 530 1320 730 270 ...
##  $ Amt_tax_paid      : int  1203 4184 4010 1819 1376 763 1438 4893 3765 1794 ...
##  $ Num_MI            : int  290 740 520 190 110 0 370 910 560 200 ...
##  $ Amt_MI            : int  2499 6343 4436 1688 1298 0 3170 7314 4707 1820 ...
##  $ Child_credit      : int  80 390 140 60 0 0 80 360 140 60 ...
##  $ Amt_child_credit  : int  28 271 108 48 0 0 32 248 102 52 ...
```

The zipcodes were changed to factors. In addition, zipcodes equal to 99999 were deleted as these are masked zipcodes used for privacy protection.

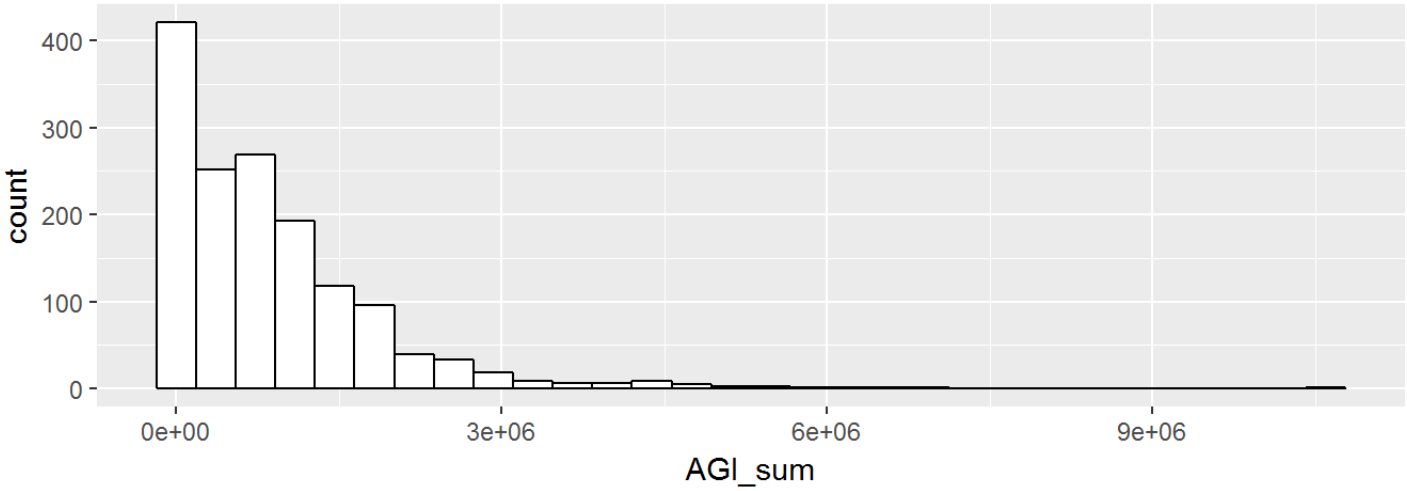
```
##  State        Zipcode     AGI_category  Num_returns    Num_single_returns
##  CA:8897   90001  :   6   1:1483       Min.   :    0   Min.   :    0.0
##            90002  :   6   2:1483       1st Qu.:  200   1st Qu.:   50.0
##            90003  :   6   3:1483       Median : 1050   Median :  290.0
##            90004  :   6   4:1483       Mean   : 1882   Mean   :  894.8
##            90005  :   6   5:1482       3rd Qu.: 2580   3rd Qu.: 1010.0
##            90006  :   6   6:1483       Max.   :24110   Max.   :13830.0
##            (Other):8861
##  Num_joint_returns Num_dependents       AGI            Num_ord_div
##  Min.   :   0.0    Min.   :    0   Min.   :       0   Min.   :   0.0
##  1st Qu.: 100.0    1st Qu.:  120   1st Qu.:   15241   1st Qu.:  40.0
##  Median : 420.0    Median :  660   Median :   68705   Median : 180.0
##  Mean   : 679.3    Mean   : 1435   Mean   :  144328   Mean   : 333.5
##  3rd Qu.: 980.0    3rd Qu.: 1820   3rd Qu.:  153623   3rd Qu.: 450.0
##  Max.   :7010.0    Max.   :28420   Max.   :10201570   Max.   :6340.0
##
##   Amt_ord_div      Num_bus_inc    Amt_bus_inc      Num_cap_gain
##  Min.   :     0   Min.   :   0   Min.   :  -152   Min.   :   0.0
##  1st Qu.:   122   1st Qu.:  40   1st Qu.:   578   1st Qu.:  40.0
##  Median :   617   Median : 190   Median :  2973   Median : 170.0
##  Mean   :  3280   Mean   : 345   Mean   :  5905   Mean   : 318.4
##  3rd Qu.:  1864   3rd Qu.: 450   3rd Qu.:  7073   3rd Qu.: 420.0
##  Max.   :361292   Max.   :6960   Max.   :187283   Max.   :6560.0
##
##   Amt_cap_gain      Num_tax_paid     Amt_tax_paid         Num_MI
##  Min.   :   -123   Min.   :   0.0   Min.   :      0   Min.   :   0.0
##  1st Qu.:    101   1st Qu.:  80.0   1st Qu.:    560   1st Qu.:  50.0
##  Median :    622   Median : 370.0   Median :   2385   Median : 240.0
##  Mean   :  10907   Mean   : 649.6   Mean   :  10940   Mean   : 464.6
##  3rd Qu.:   2170   3rd Qu.: 890.0   3rd Qu.:   7246   3rd Qu.: 620.0
##  Max.   :7061052   Max.   :7650.0   Max.   :1300646   Max.   :6240.0
##
##      Amt_MI        Child_credit     Amt_child_credit
##  Min.   :     0   Min.   :   0.00   Min.   :  0.00
##  1st Qu.:   506   1st Qu.:   0.00   1st Qu.:  0.00
##  Median :  2442   Median :  40.00   Median : 18.00
##  Mean   :  5755   Mean   :  74.57   Mean   : 42.67
##  3rd Qu.:  6697   3rd Qu.: 110.00   3rd Qu.: 60.00
##  Max.   :104486   Max.   :1350.00   Max.   :722.00
##
```

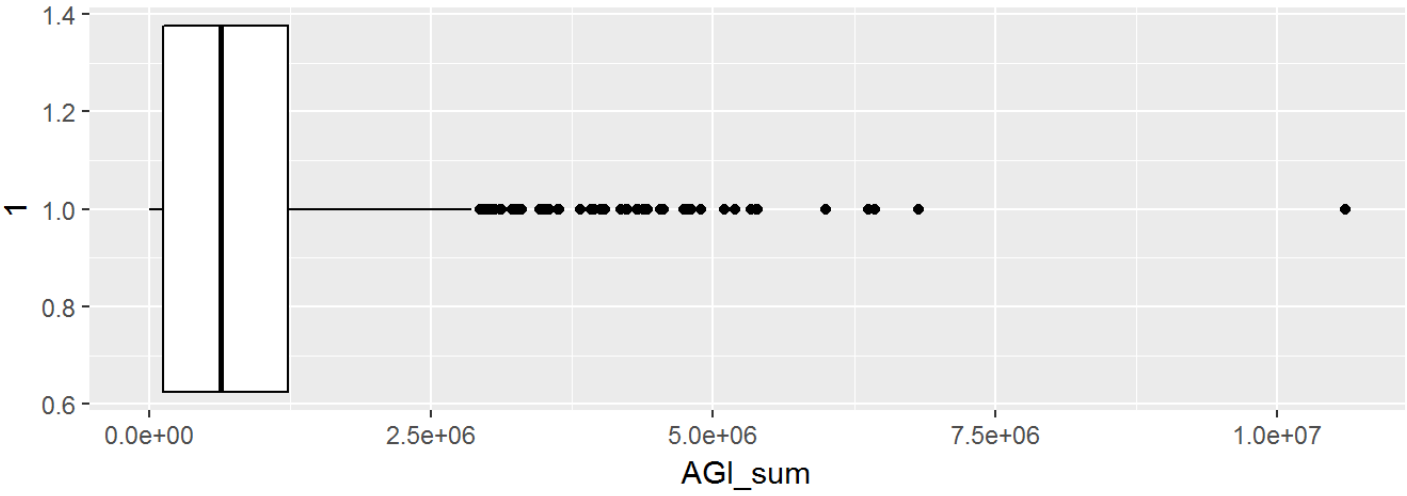
There are 8,903 observations of 20 variables in this data set.

The AGI_category variable ranges from 1-6 as defined below based on the adjusted gross income of the tax return: 1 = 1 - 25,000 2 = 25,000 - 50,000 3 = 50,000 - 75,000 4 = 75,000 - 100,000 5 = 100,000 - 200,000 6 = 200,000 or more

In order to examine data for each zipcode, the data for each AGI category was added for each zipcode. This new dataset was then explored.

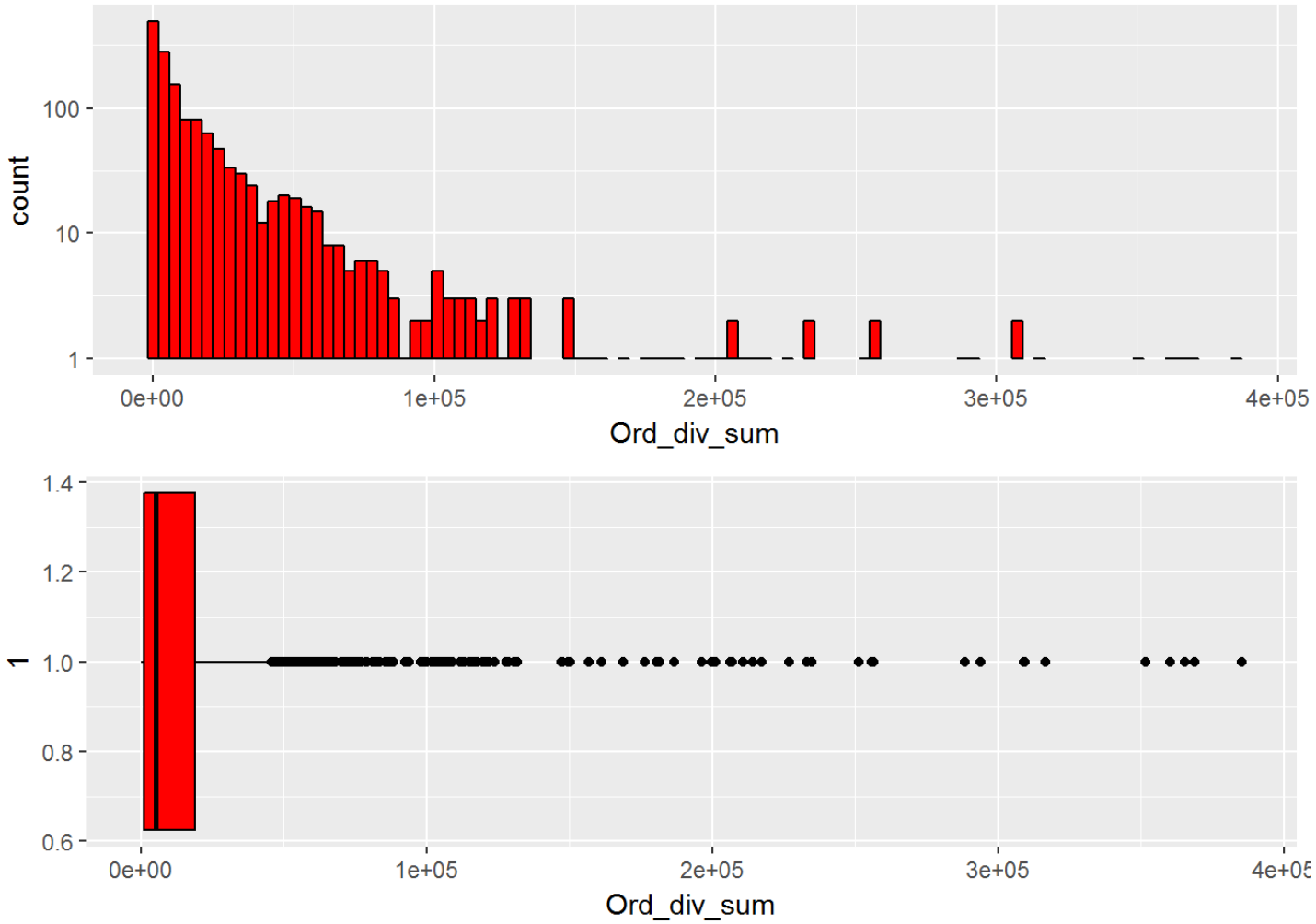
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1483 obs. of  8 variables:
##  $ Zipcode     : Factor w/ 1483 levels "90001","90002",..: 1 2 3 4 5 6 7 8 9 10 ..
.
##  $ AGI_sum     : int  551699 488196 640364 1714986 753818 604508 328241 688854 461
580 888835 ...
##  $ Ord_div_sum : int  197 156 101 44726 7168 1152 908 3792 15977 232 ...
##  $ Bus_inc_sum : int  40935 42228 59935 102340 47257 54465 22532 24678 26134 73075
...
##  $ Cap_gain_sum: int  760 132 306 97731 41877 10449 3008 14770 51568 2229 ...
##  $ Tax_paid_sum: int  13355 13227 14201 124143 44454 11764 7634 39110 51642 16633
...
##  $ MI_sum      : int  16264 18014 20414 42312 14803 10249 6409 39335 10063 23478 .
..
##  $ Child_sum   : int  455 434 519 381 231 263 165 333 29 598 ...
```





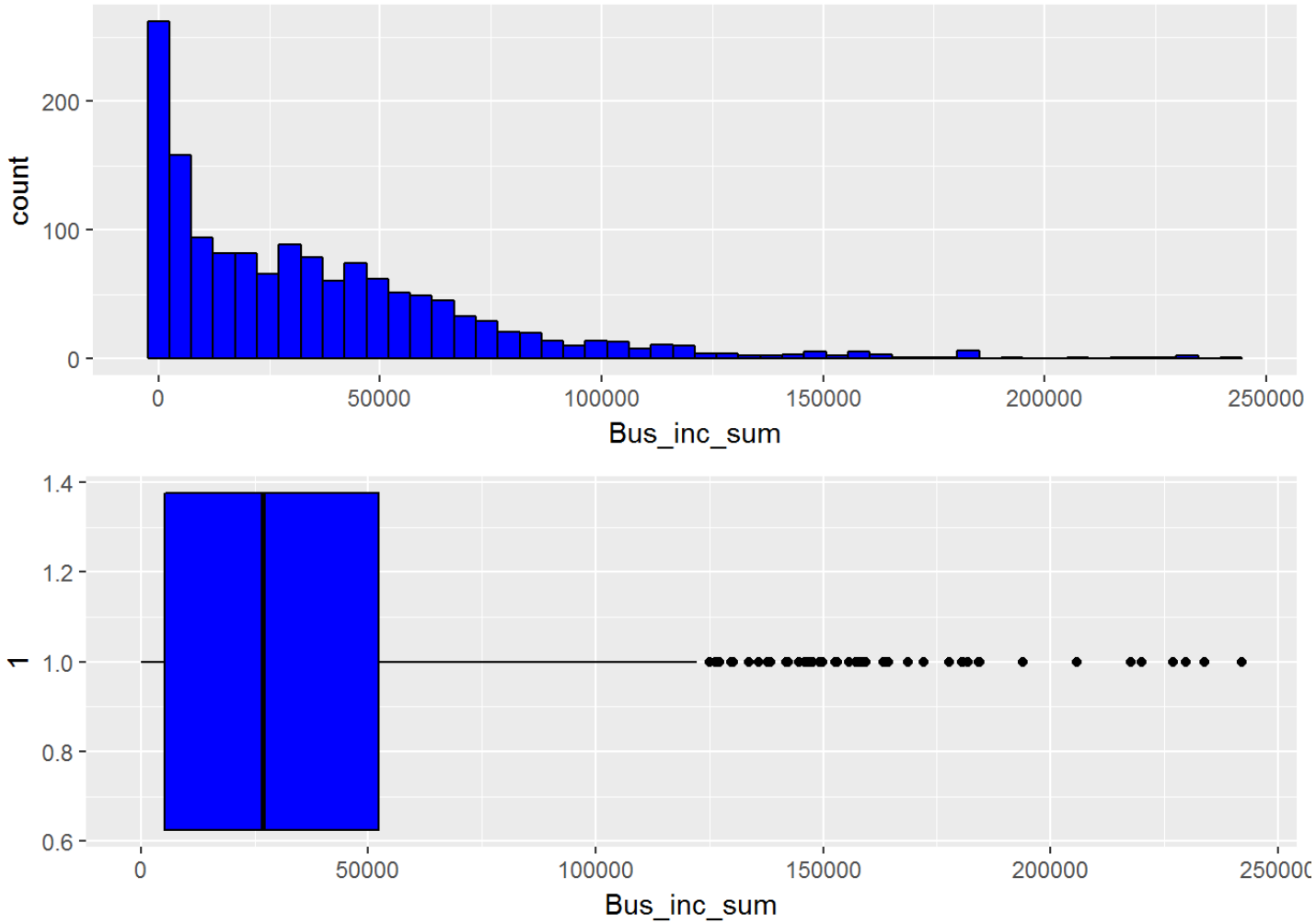

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     3155   128900   636800   865900  1230000 10600000
```

The average AGI per zipcode was 865,900. There is a wide range of values from 3,155 to 10,600,000. The distribution is positively skewed. Most of the data is below 3,000,000 as shown in the box plot.



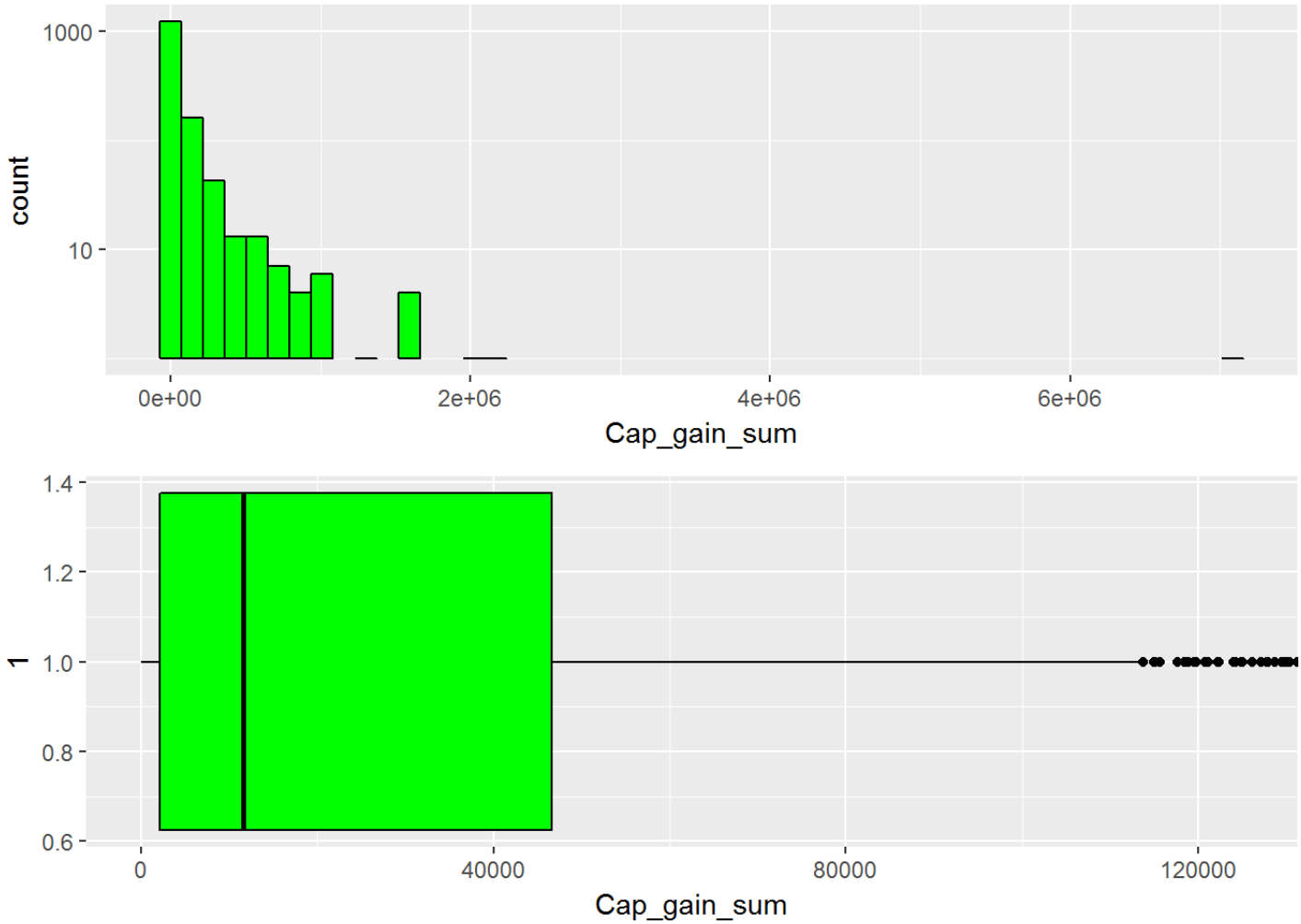

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0    975.5   5280.0  19680.0  18880.0 385300.0
```

The average ordinary dividends reported was 5,280 with a maximum of 385,300. A log10 transformation was applied to the y axis. The distribution is positively skewed. There are several gaps in the data after at values greater than 100,000. Most of the data is below 50,000 excluding outliers.



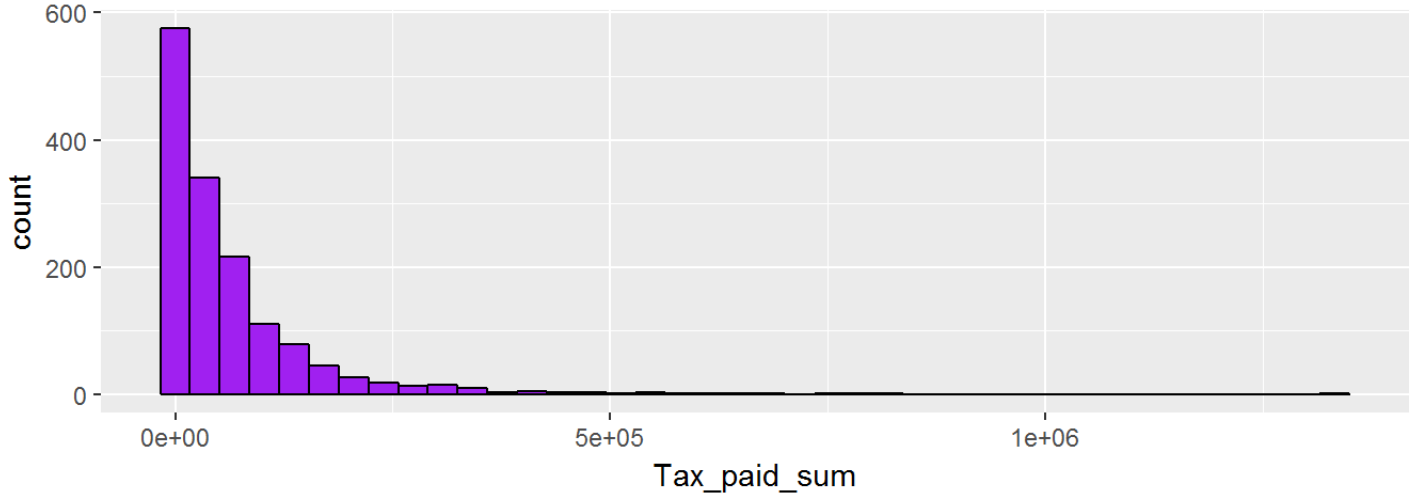

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    5138   26850   35430   52250  242100
```

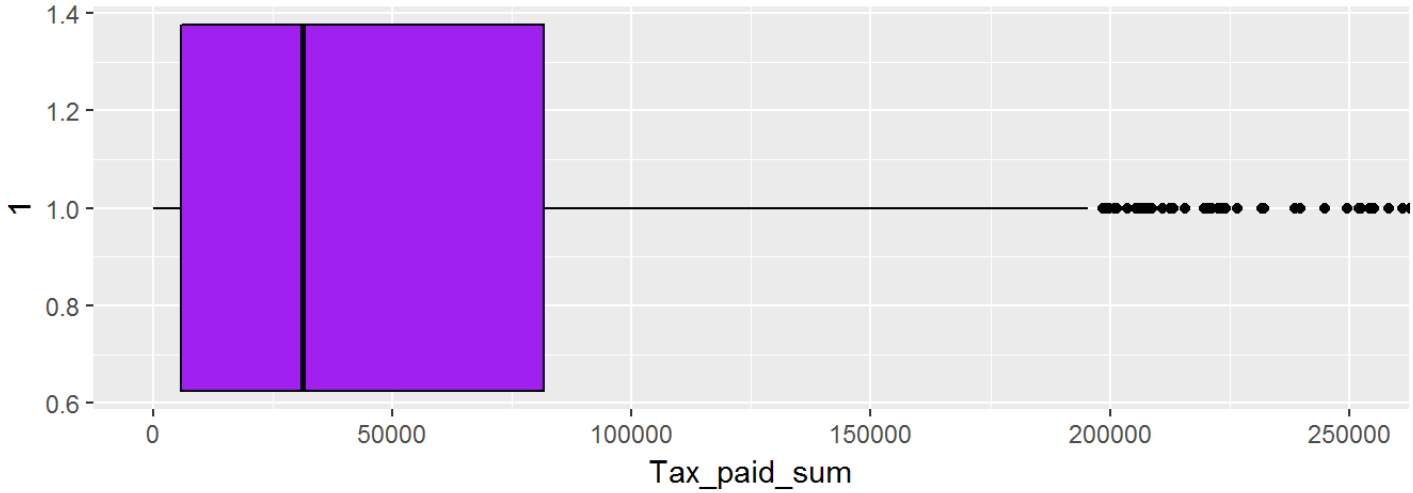
The average business income reported was 35,430 with a maximum of 242,100. The distribution is positively skewed. Most of the data is below 100,000 as shown in the box plot.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     -20    2074   11580   65430   46630 7083000
```

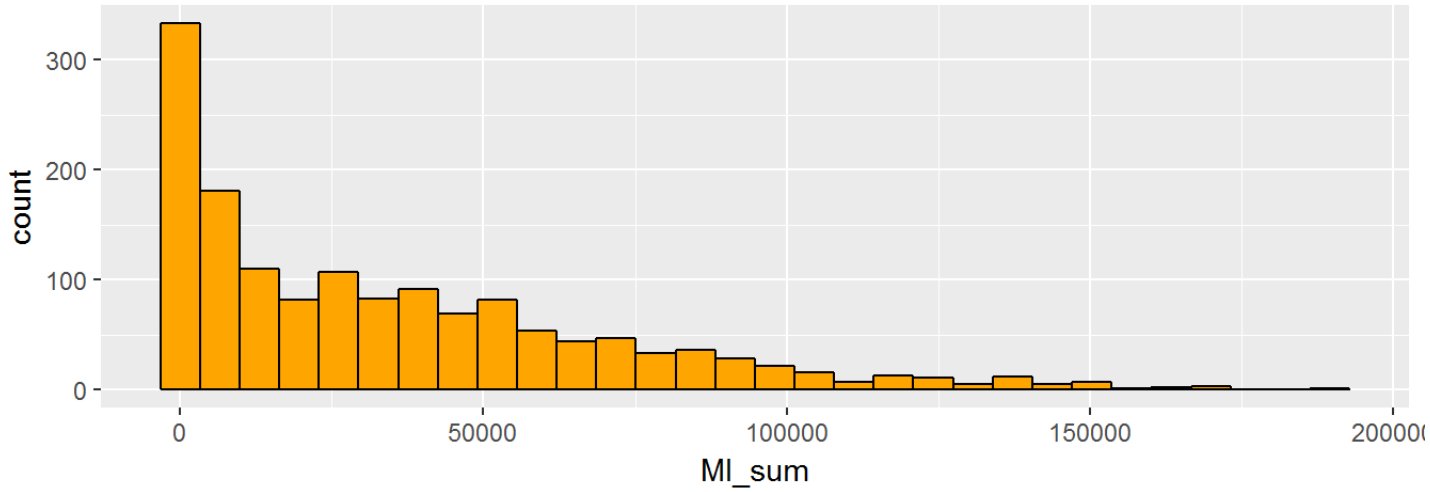
The average net capital gains was 65,430 with a positively skewed distribution. The negative minimum value of -20 denotes greater capital loss than gain. There is a gap in the distribution from about 1,000,000 to 2,000,000. Data below about 120,000 excludes outliers.

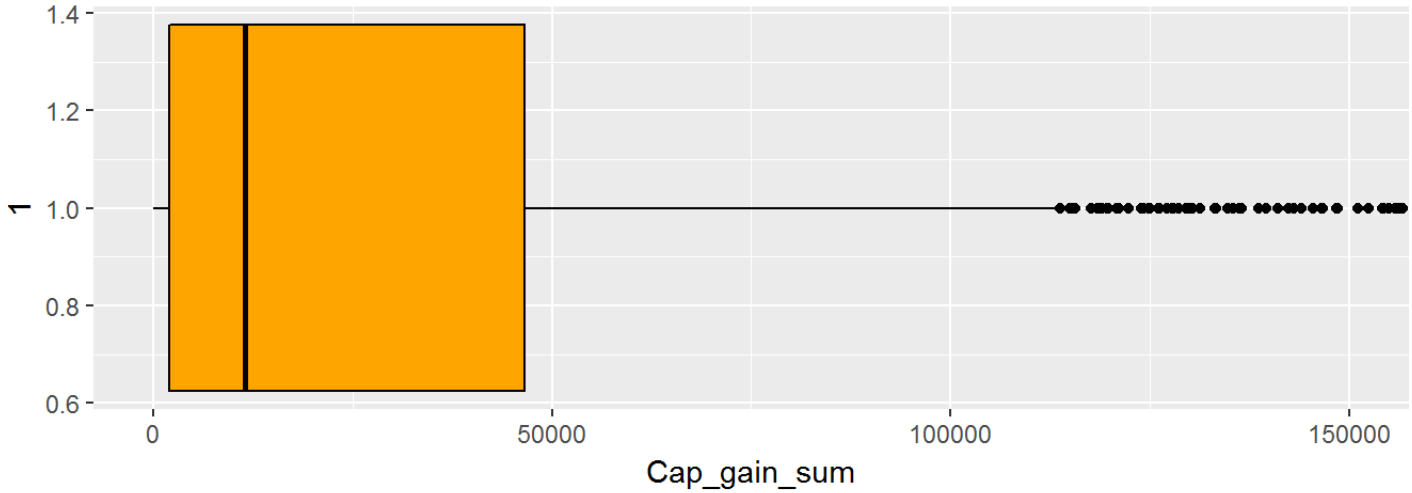





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    5808   31200   65640   81760 1332000
```

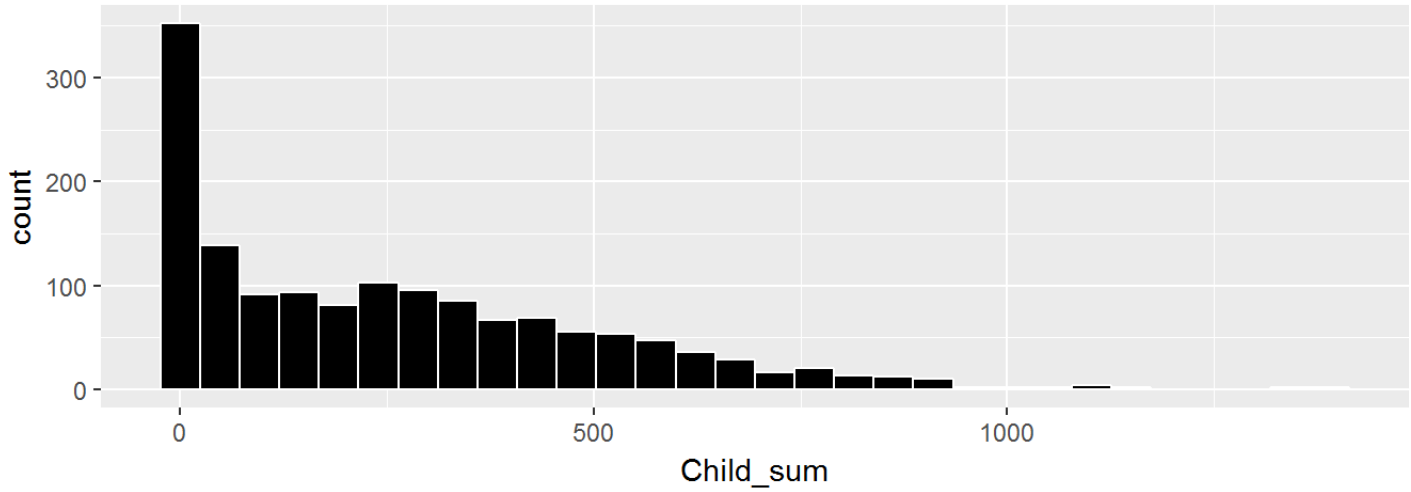
The average amount of taxes paid per zipcode was 65,640. The distribution is positively skewed, which is similar to the previous plots. Outliers are greater than 200,000.

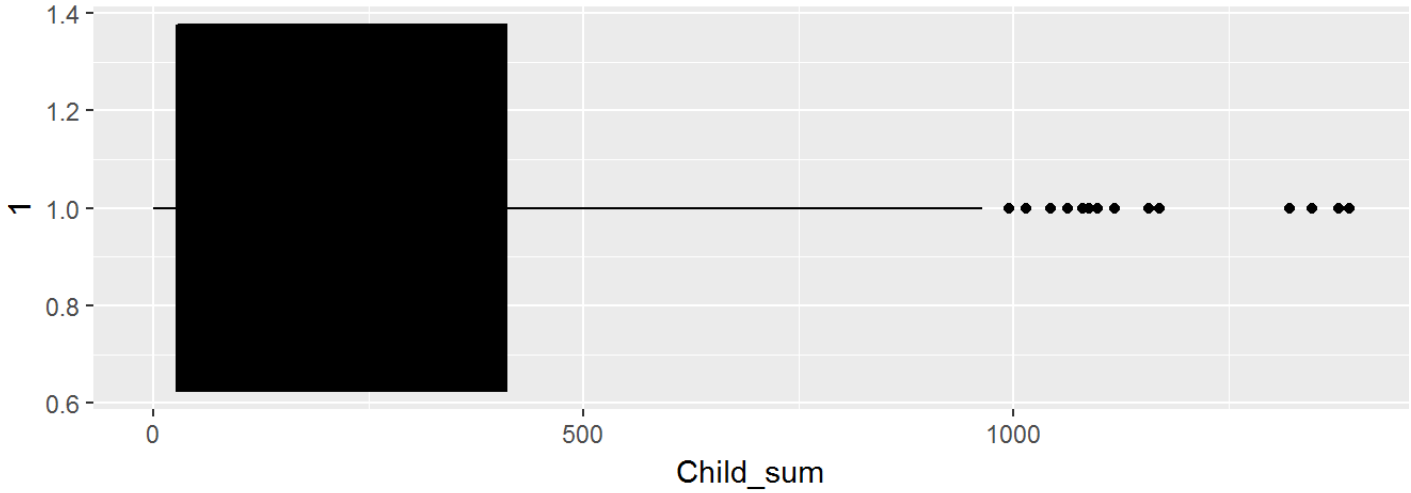





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    4346   25070   34530   53030  189500
```

The average mortgage interest deduction was 34,530 with most of the data below 125,000 excluding outliers.






```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    27.5   209.0   256.0   410.5  1390.0
```

The average child and dependent care deduction was 256. Most of the data is below 1,000 excluding outliers.

Univariate Analysis

What is the structure of your dataset?

There are 1,483 observations of 9 variables.

What is/are the main feature(s) of interest in your dataset?

The main feature is the AGI. The correlation between other variables and AGI will be explored further.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

Business income (Bus_inc_sum) and net capital gains (Cap_gain_sum) may correlate with AGI.

Did you create any new variables from existing variables in the dataset?

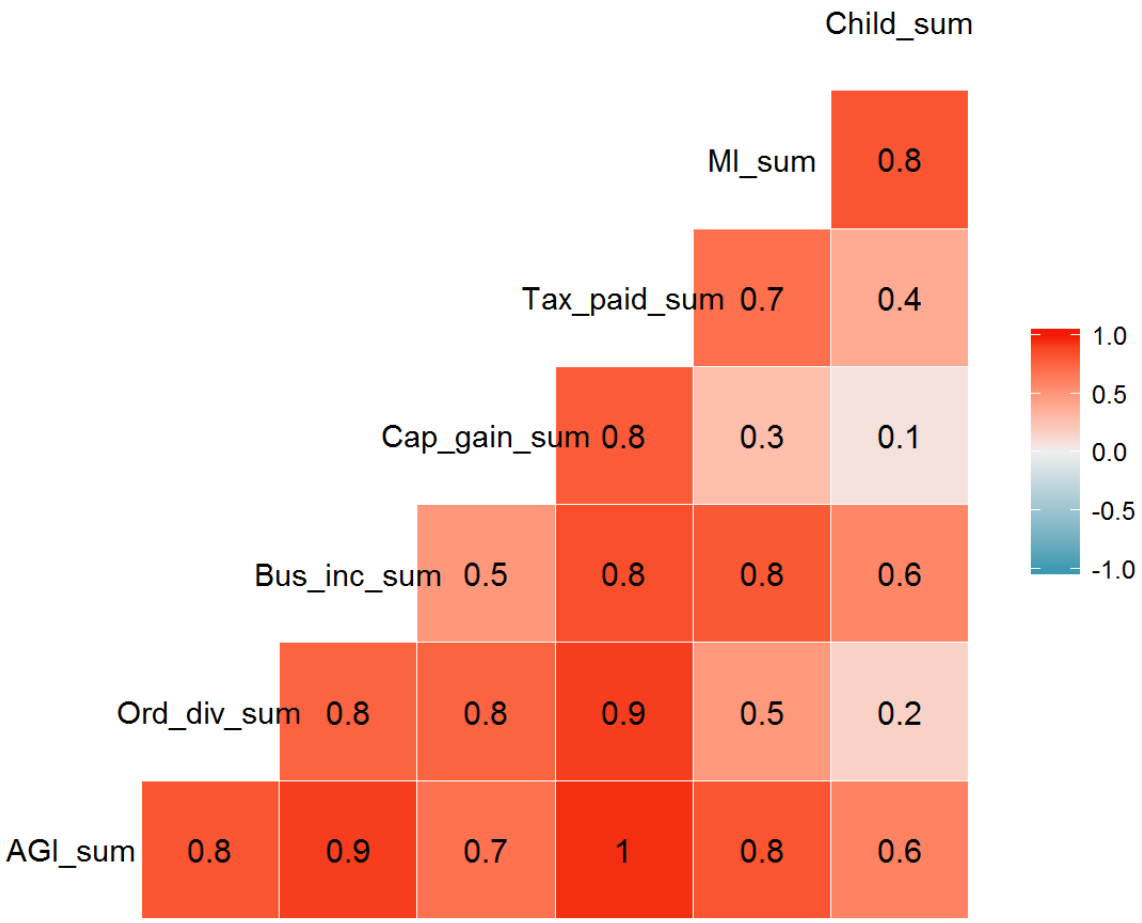
The new variables of AGI_sum, Ord_div_sum, Bus_inc_sum, Cap_gain_sum, Tax_paid_sum, MI_sum and Child_sum were created from the original dataset. These were created by adding the data for each zipcode across AGI categories.

Of the features you investigated, were there any unusual distributions?
Did you perform any operations on the data to tidy, adjust, or change the form
of the data? If so, why did you do this?

The Ord_div_sum and Cap_gain_sum were log10 transformed on the y axis of the historgram due to the wide range in counts. The transformed distribution of both variables contained gaps.

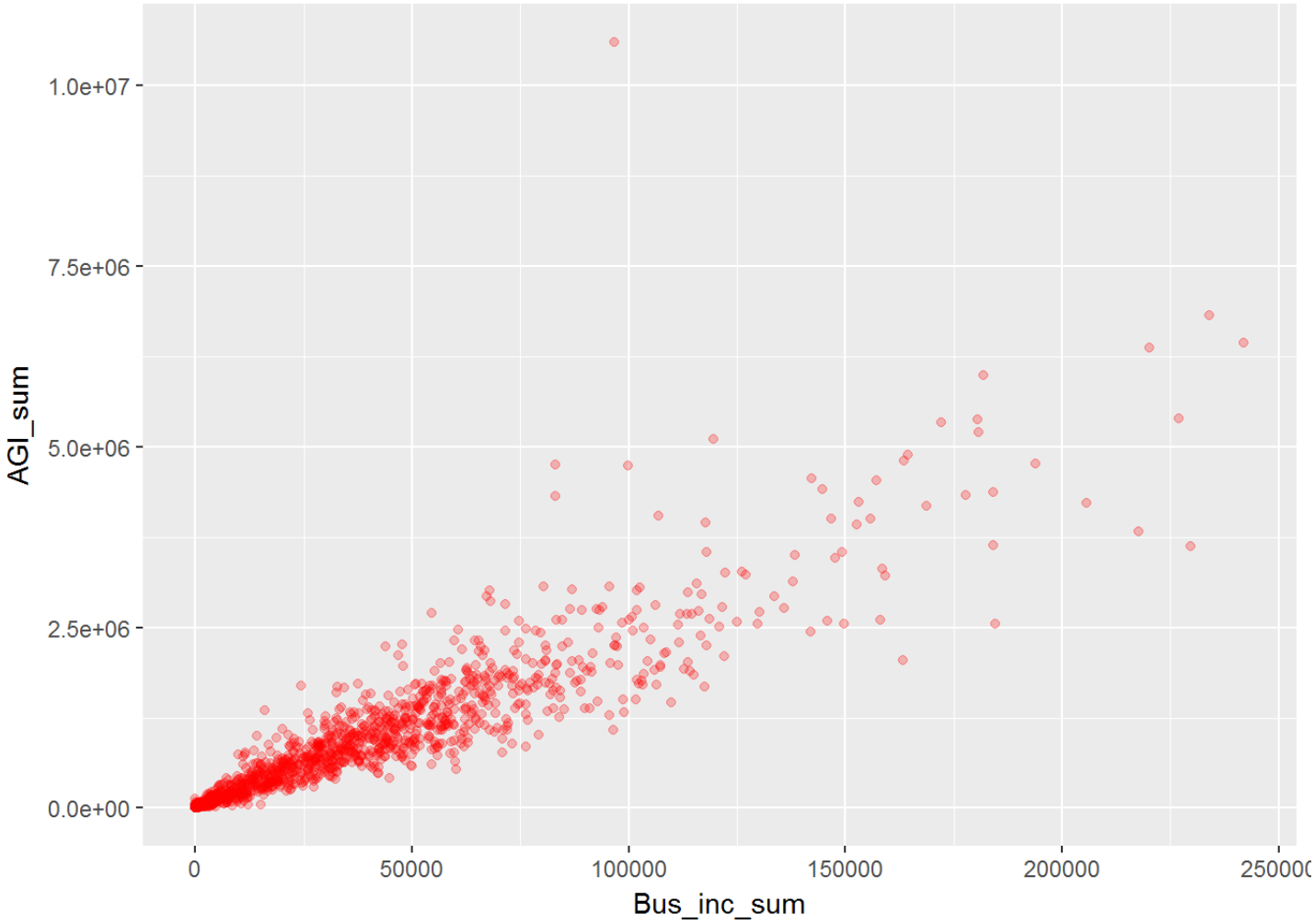
Bivariate Plots Section

Next, a correlation plot of each variable in the new dataset was plotted.

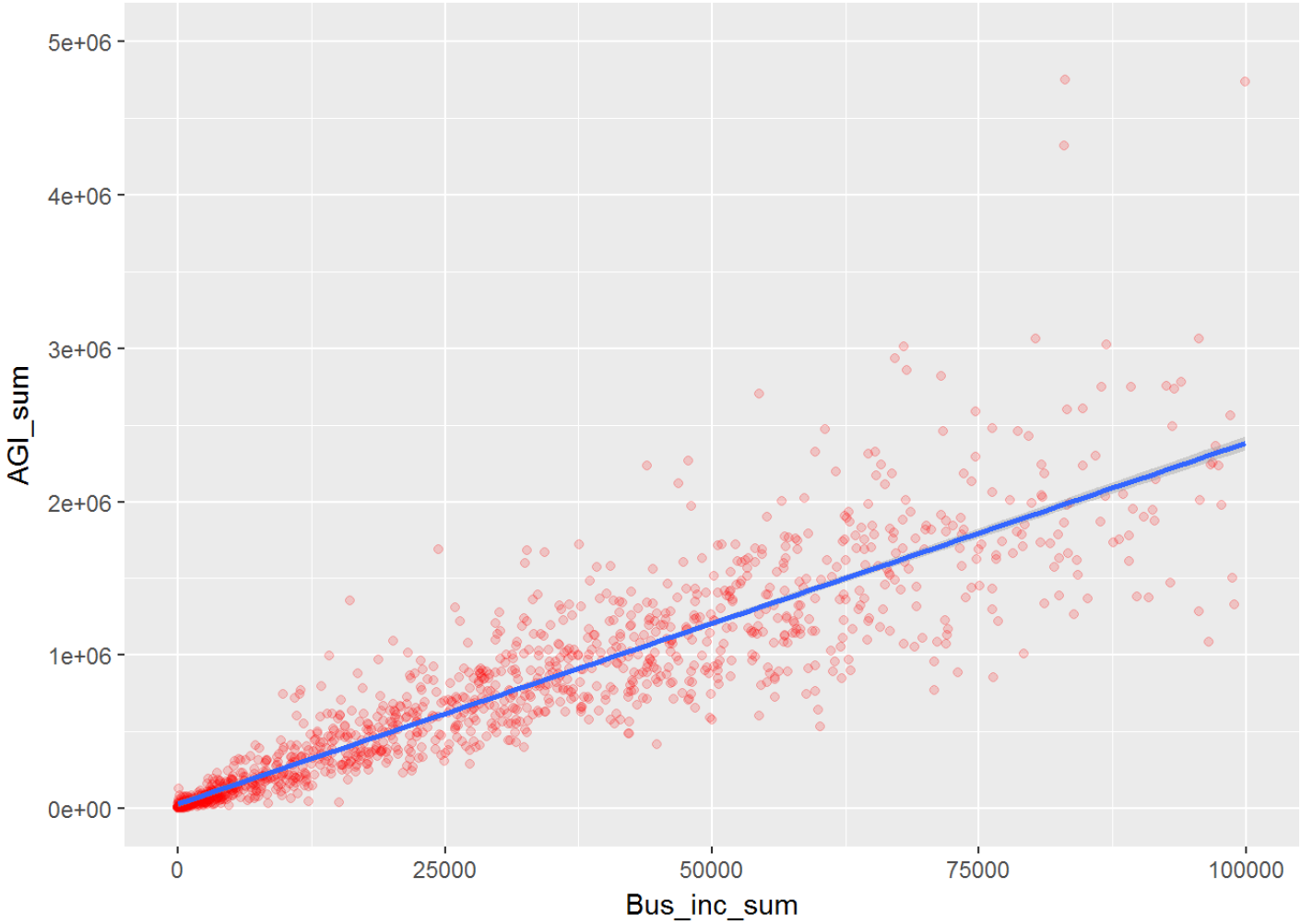


From this correlation plot, the following variables showed correlation and will be explored further:

1. AGI (AGI_sum) with business net income (Bus_inc_sum)
2. AGI with mortage interest amount (MI_sum)
3. Child and dependent care credit amount (Child_sum) with MI_sum
4. Business net income with ordinary dividends amount (Ord_div_sum)

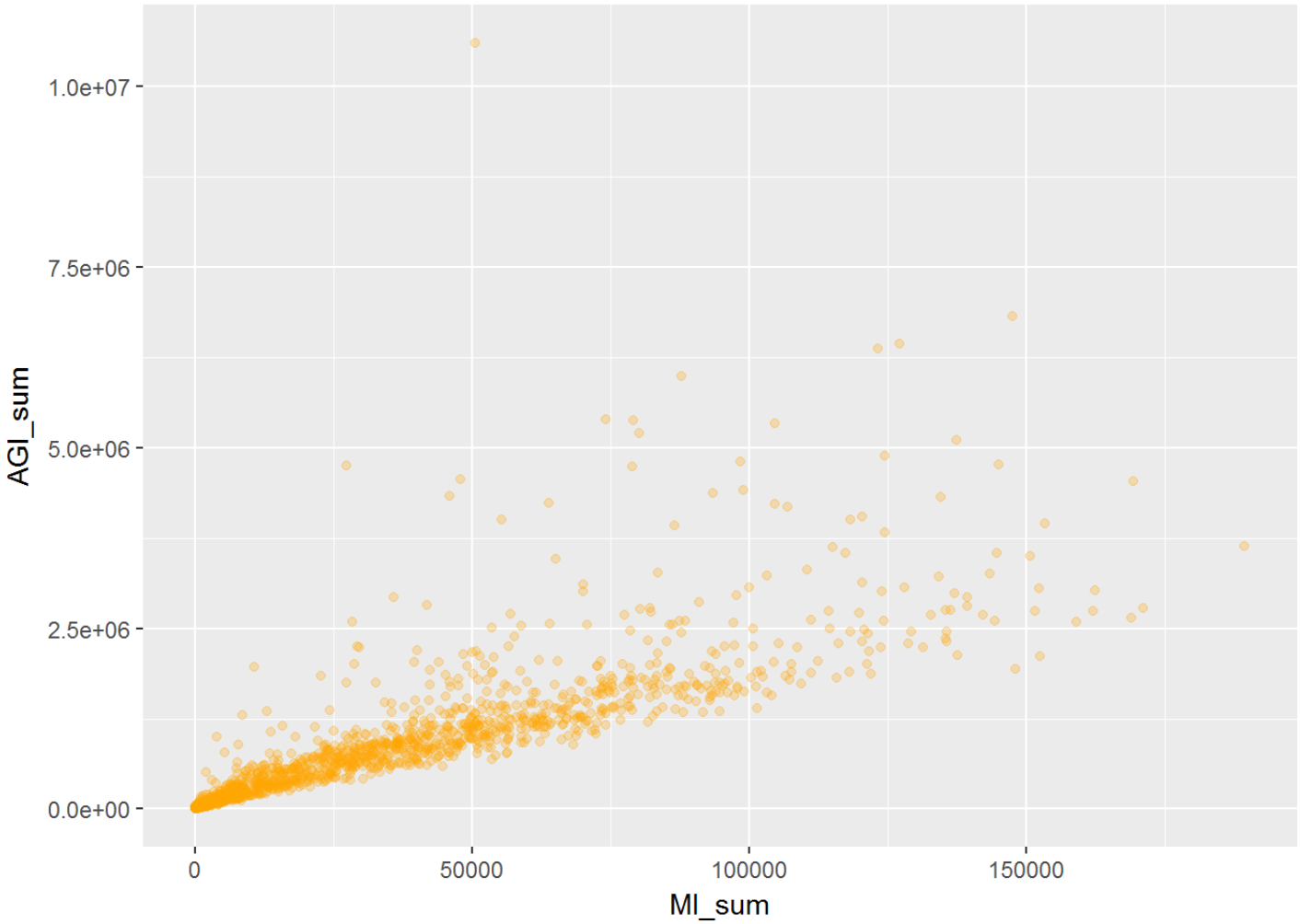


This plot shows that the data is concentrated in the range of 0 to 100,000 for Bus_inc_sum. The next plot shows the data in this range.

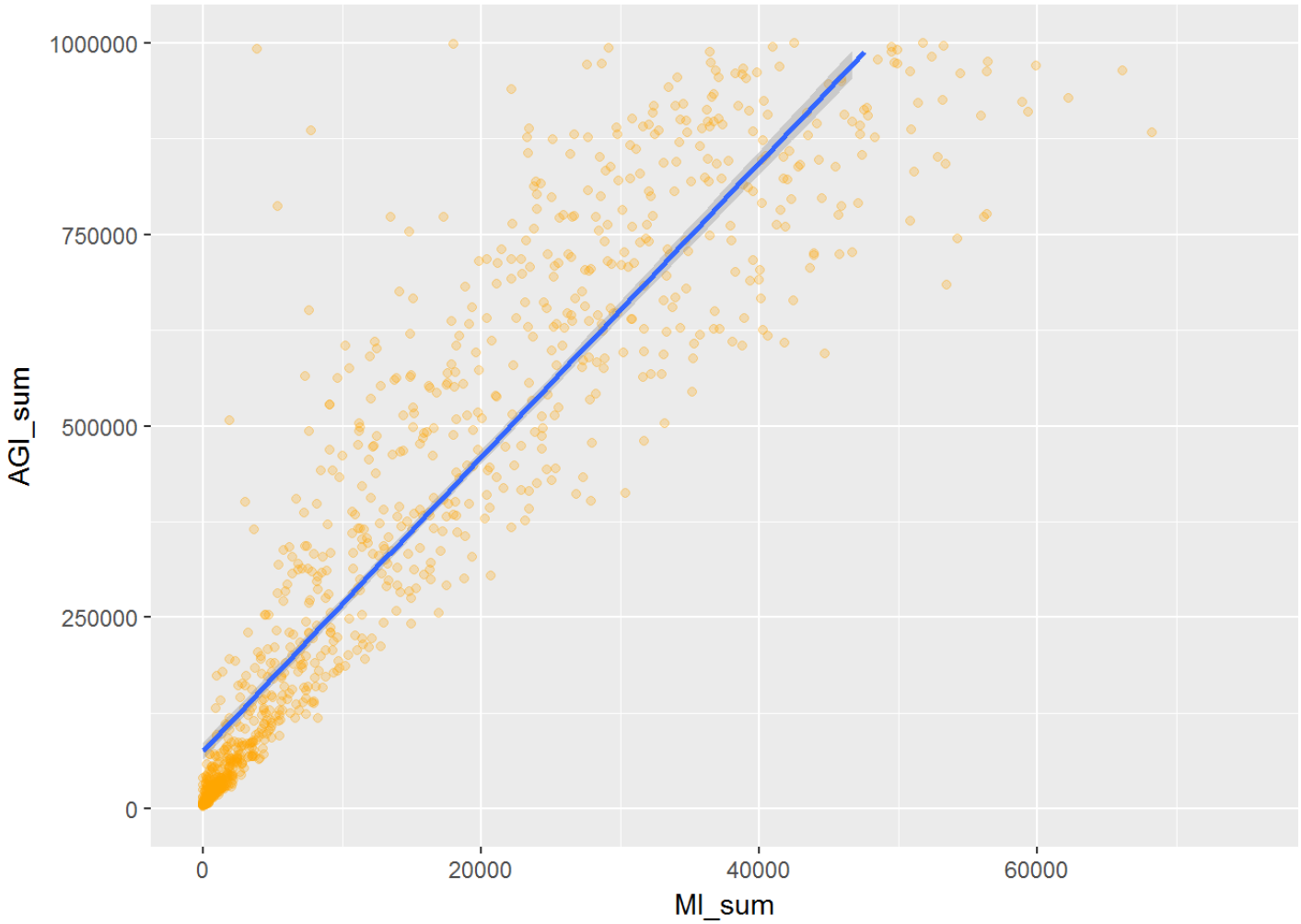



```
##
##  Pearson's product-moment correlation
##
## data:  taxes.zipcode$AGI_sum and taxes.zipcode$Bus_inc_sum
## t = 83.883, df = 1481, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.899631 0.917371
## sample estimates:
##       cor
## 0.9089114
```

This plot shows data closer to the linear model fit line in the range of about 0 to 75,000 of Bus_inc_sum with more disparate data for higher values. The correlation coefficent is 0.91, which implies that these two variables are closely correlated. Increasing business income is correlated with increasing AGI for zipcode areas.

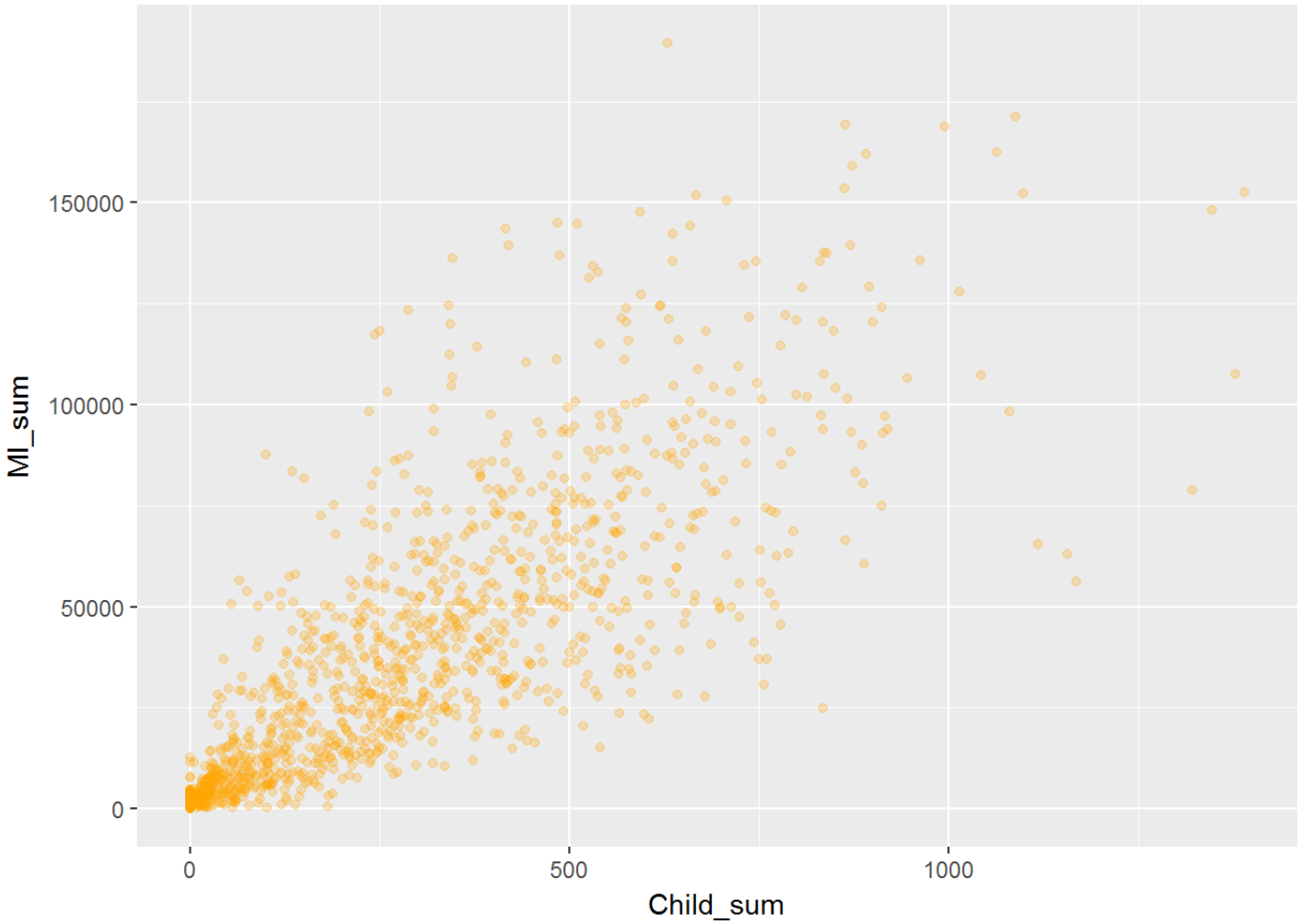


This plot shows that the data is concentrated at lower values of MI_sum and AGI_sum. The x and y limits will be adjusted in the next plot to exclude outliers.

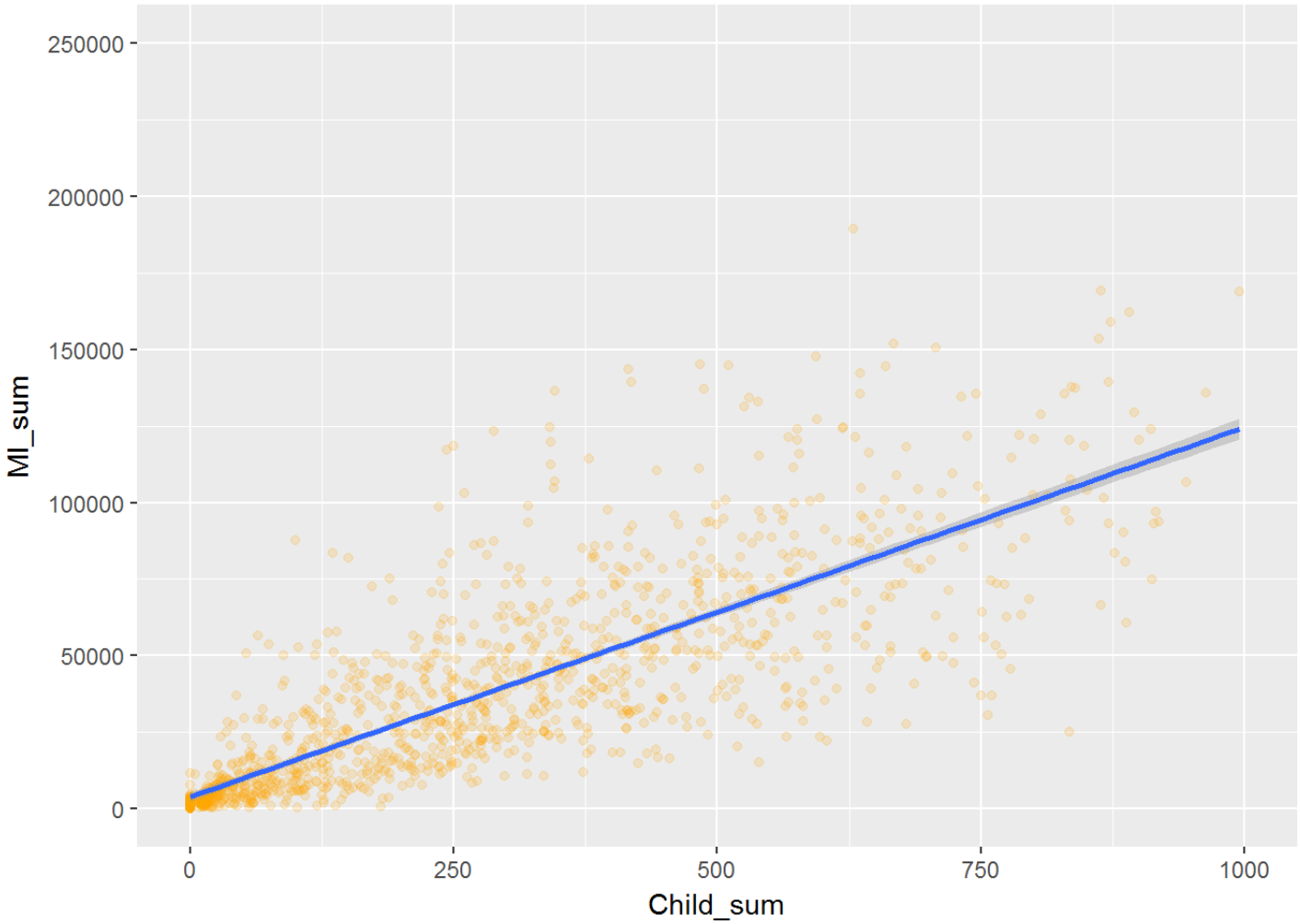



```
##
##  Pearson's product-moment correlation
##
## data:  taxes.zipcode$AGI_sum and taxes.zipcode$MI_sum
## t = 56.432, df = 1481, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8093071 0.8416810
## sample estimates:
##       cor
## 0.8261748
```

The data closely fits the linear fit line in the MI_sum range of 0 to 10,000. There is more variation from the linear fit line at higher values. The correlation coefficient was 0.83, which shows that the mortage interest deduction and AGI are correlated.

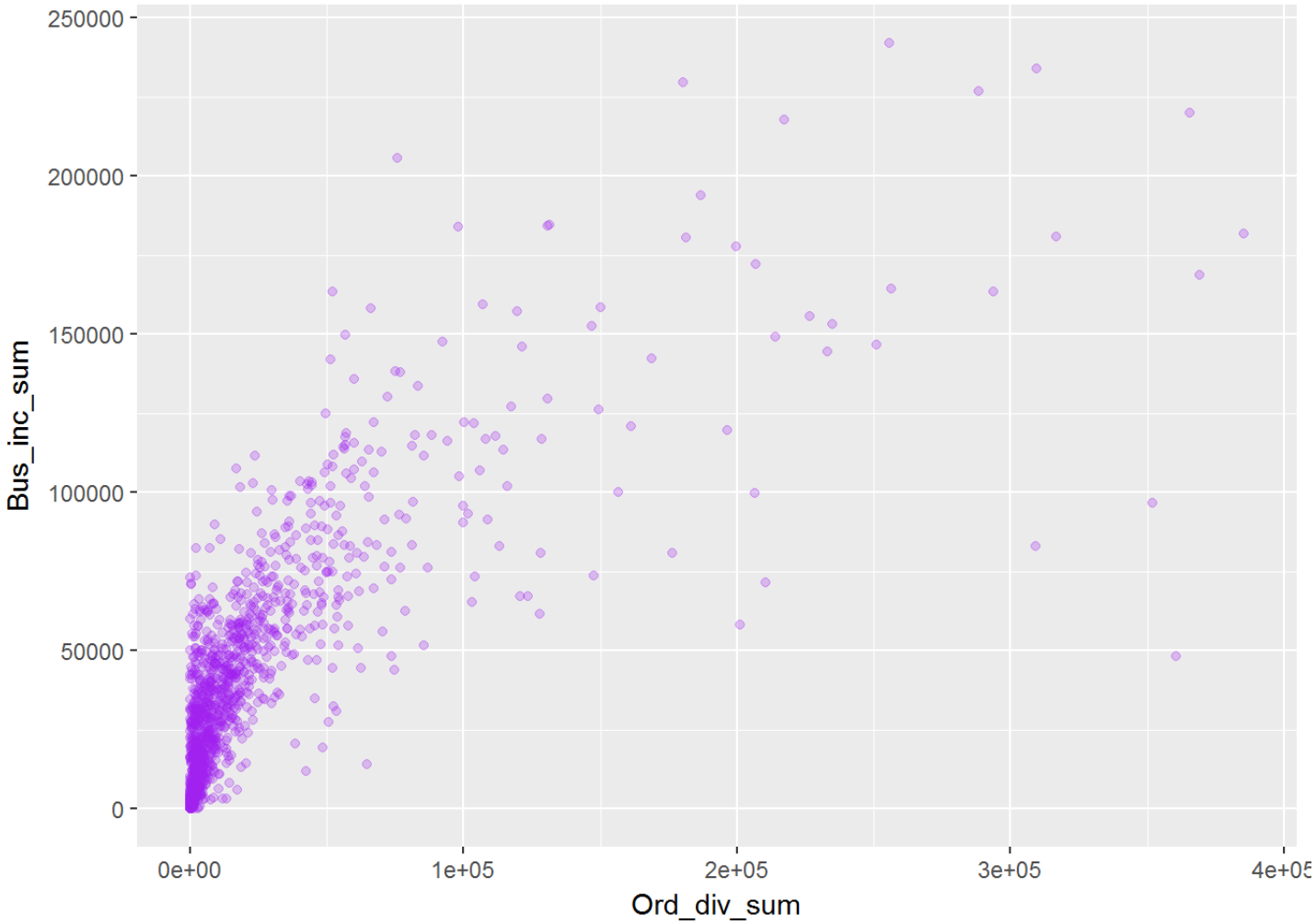


This plot shows the data is concentrated in the Child_sum range of 0 to 1000. The next plot will exclude outliers.

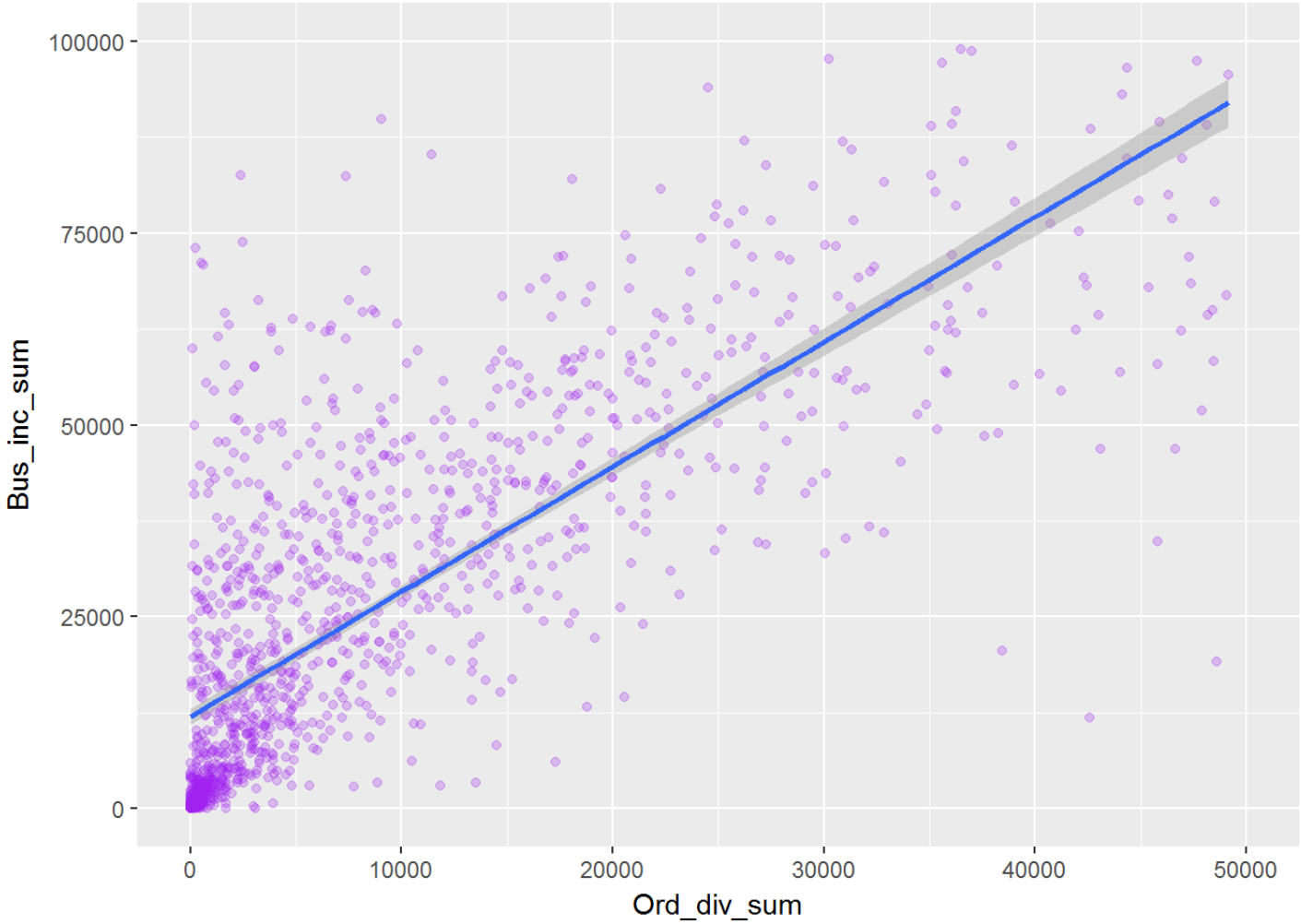



```
##
##  Pearson's product-moment correlation
##
## data:  taxes.zipcode$Child_sum and taxes.zipcode$MI_sum
## t = 55.244, df = 1481, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8031794 0.8364996
## sample estimates:
##       cor
## 0.8205354
```

This plot shows that the data better fits the linear model in the Child_sum range of 0 to 250. At higher values, the data varies more with the linear model. The correlation coefficient was 0.82, showing correlation between the child and dependent care tax deduction and mortgage interest deduction.



This scatterplot of Bus_inc_sum vs. Ord_div_sum shows that the data is concentrated in the 0 to 50,000 range of Ord_div_sum. The next plot will focus on this range.




```
##
##  Pearson's product-moment correlation
##
## data:  taxes.zipcode$Ord_div_sum and taxes.zipcode$Bus_inc_sum
## t = 44.245, df = 1481, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7317242 0.7756361
## sample estimates:
##       cor
## 0.7545234
```

This plot shows the relationship between Ord_div_sum and Bus_inc_sum and fits the data to a linear model line. The data shows greater variation with the linear model at Ord_div_sum values greater than 5,000. The correlation coefficient was 0.75, showing correlation between the two variables.

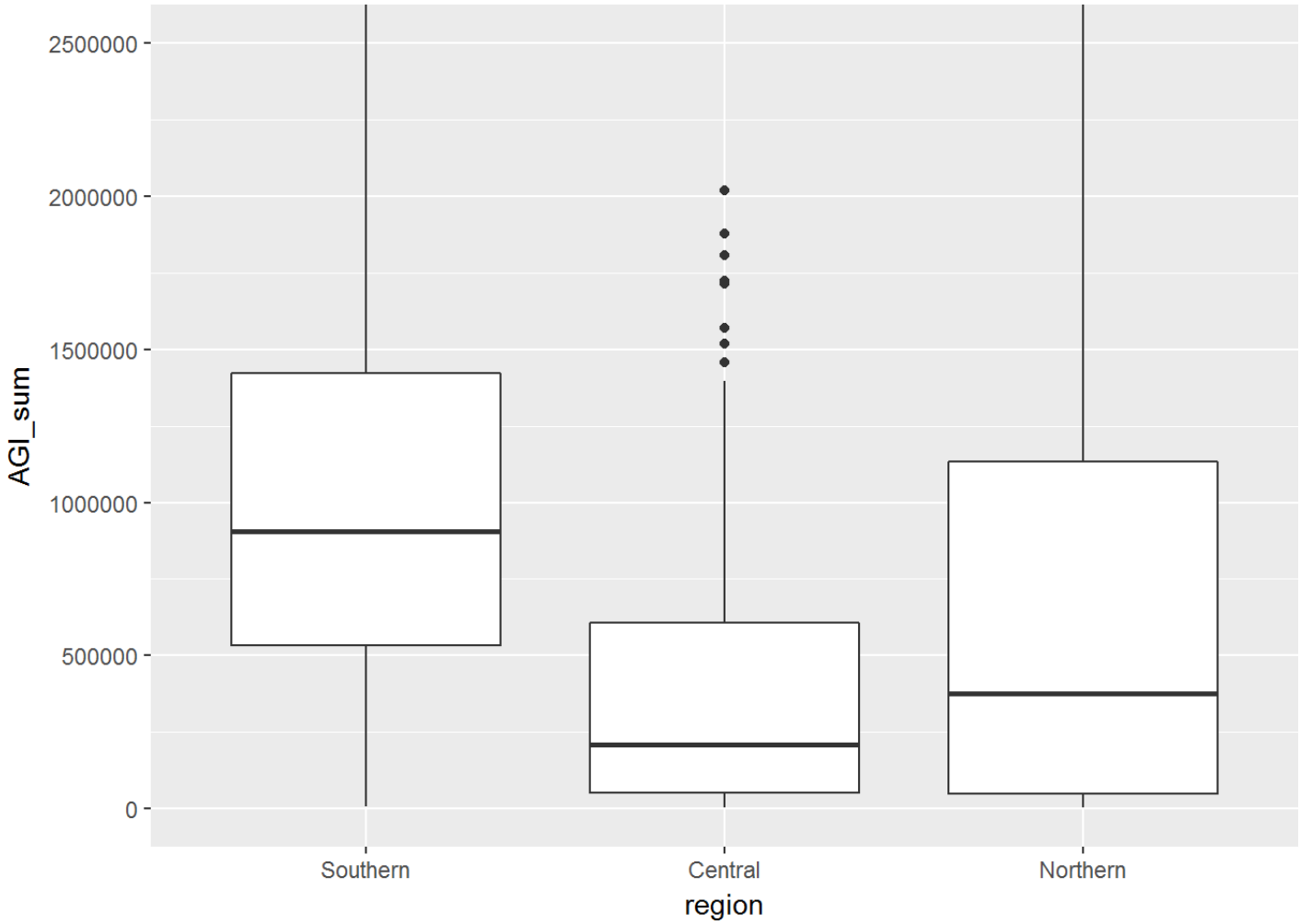
Next, the regions of Northern, Central and Southern California were grouped by using zipcode, county and map references. The data for these regions was then analyzed.

The zipcode variable was first cut and a new region variable was created. The zipcode ranges correspond to zipcode ranges in Northern, Central and Southern California.

```
##     Zipcode         AGI_sum          Ord_div_sum        Bus_inc_sum
##  Min.   :90001   Min.   :    3155   Min.   :     0.0   Min.   :     0
##  1st Qu.:92122   1st Qu.:  128887   1st Qu.:   975.5   1st Qu.:  5138
##  Median :93610   Median :  636777   Median :  5280.0   Median : 26846
##  Mean   :93535   Mean   :  865868   Mean   : 19680.3   Mean   : 35428
##  3rd Qu.:95311   3rd Qu.: 1229637   3rd Qu.: 18881.0   3rd Qu.: 52250
##  Max.   :96161   Max.   :10602300   Max.   :385317.0   Max.   :242051
##   Cap_gain_sum      Tax_paid_sum         MI_sum         Child_sum
##  Min.   :    -20   Min.   :      0   Min.   :     0   Min.   :   0.0
##  1st Qu.:   2074   1st Qu.:   5808   1st Qu.:  4346   1st Qu.:  27.5
##  Median :  11584   Median :  31203   Median : 25069   Median : 209.0
##  Mean   :  65432   Mean   :  65636   Mean   : 34527   Mean   : 256.0
##  3rd Qu.:  46632   3rd Qu.:  81755   3rd Qu.: 53026   3rd Qu.: 410.5
##  Max.   :7082756   Max.   :1332325   Max.   :189501   Max.   :1390.0
##                region
##  (9e+04,9.32e+04]  :617
##  (9.32e+04,9.4e+04]:201
##  (9.4e+04,9.62e+04]:665
##
##
##
```

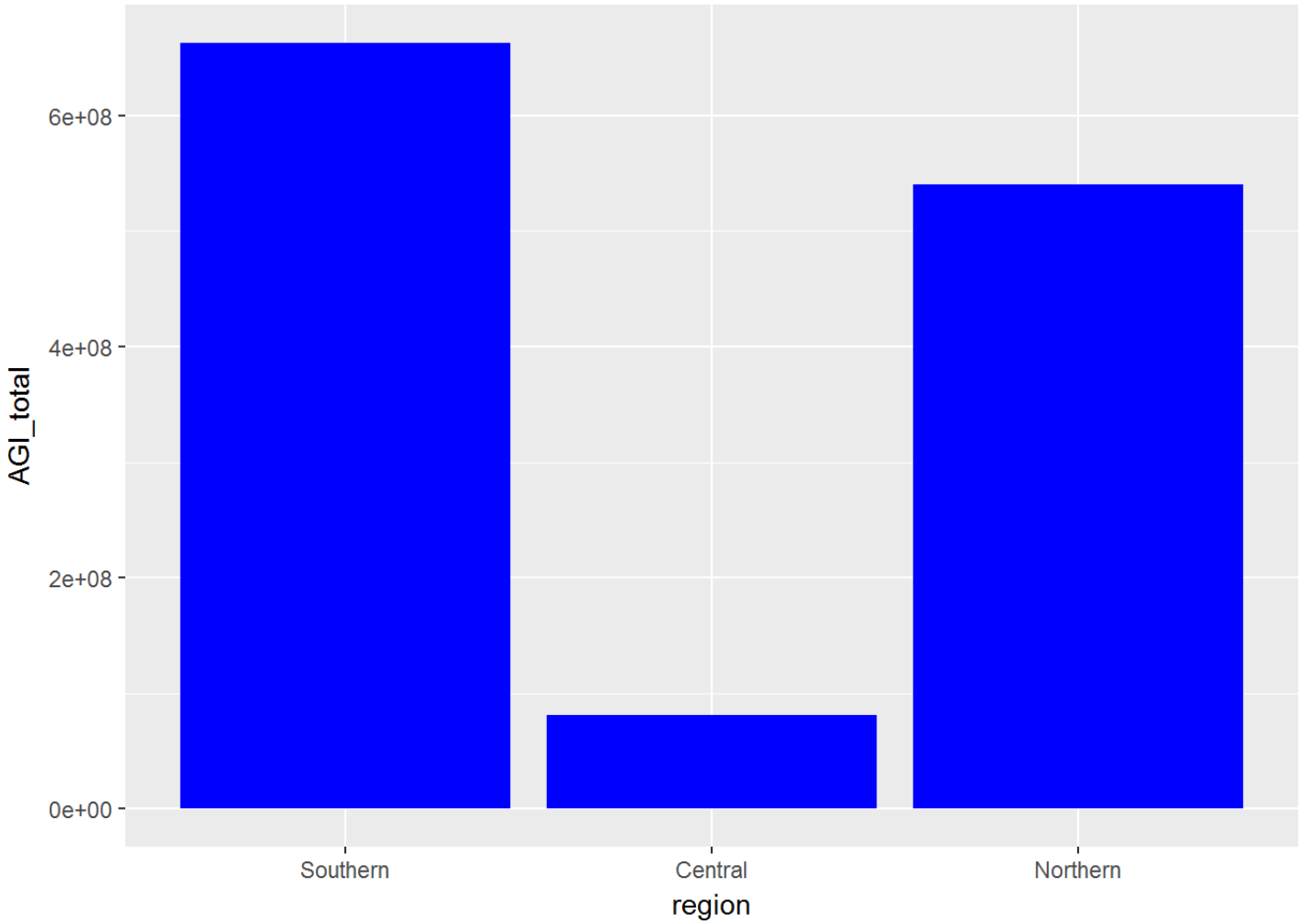
The region variable levels were then renamed for each region.

```
##     Zipcode         AGI_sum          Ord_div_sum        Bus_inc_sum
##  Min.   :90001   Min.   :    3155   Min.   :     0.0   Min.   :     0
##  1st Qu.:92122   1st Qu.:  128887   1st Qu.:   975.5   1st Qu.:  5138
##  Median :93610   Median :  636777   Median :  5280.0   Median : 26846
##  Mean   :93535   Mean   :  865868   Mean   : 19680.3   Mean   : 35428
##  3rd Qu.:95311   3rd Qu.: 1229637   3rd Qu.: 18881.0   3rd Qu.: 52250
##  Max.   :96161   Max.   :10602300   Max.   :385317.0   Max.   :242051
##   Cap_gain_sum      Tax_paid_sum         MI_sum         Child_sum
##  Min.   :    -20   Min.   :      0   Min.   :     0   Min.   :   0.0
##  1st Qu.:   2074   1st Qu.:   5808   1st Qu.:  4346   1st Qu.:  27.5
##  Median :  11584   Median :  31203   Median : 25069   Median : 209.0
##  Mean   :  65432   Mean   :  65636   Mean   : 34527   Mean   : 256.0
##  3rd Qu.:  46632   3rd Qu.:  81755   3rd Qu.: 53026   3rd Qu.: 410.5
##  Max.   :7082756   Max.   :1332325   Max.   :189501   Max.   :1390.0
##       region
##  Southern:617
##  Central :201
##  Northern:665
##
##
##
```

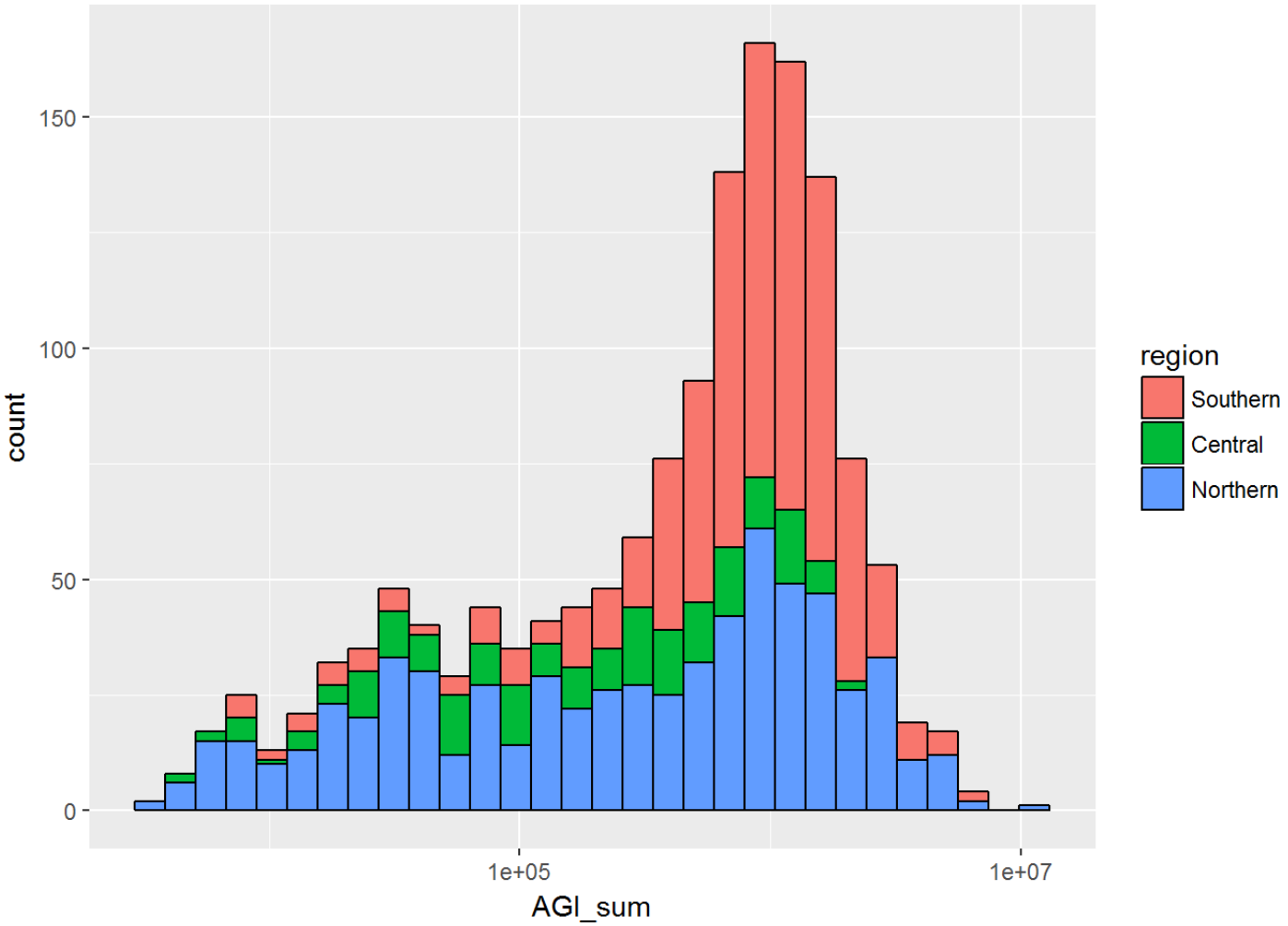



```
## subset(taxes.zipcode, !is.na(region))$region: Southern
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6890  532900  904600 1074000 1423000 6373000
## ------------------------------------------------------------
## subset(taxes.zipcode, !is.na(region))$region: Central
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4390   51690  207800  400700  607000 2020000
## ------------------------------------------------------------
## subset(taxes.zipcode, !is.na(region))$region: Northern
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     3155    47340   374400   813000  1133000 10600000
```

This boxplot shows the distribution of AGI_sum per region excluding outliers. The median AGI_sum is highest for Southern California (904,600) and lowest for Central California (374,400).



This bar graph shows the total AGI per region. The total AGI is highest for Southern California and lowest for Central California.



This plot shows the distribution of Northern California AGI_sum. More Northern California zipcodes have AGI less than 100,000 compared with Southern California. Central California values have more uniform representation in the range of AGI.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

The AGI was positively correlated with business income and mortgage interest deduction. The correlation coefficient with business income was 0.91 which shows high correlation. The correlation coefficient with mortgage interest deduction was 0.83 which shows a strong correlation.

The AGI also varied with region, with Southern California having the highest total AGI and Central California having the lowest.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

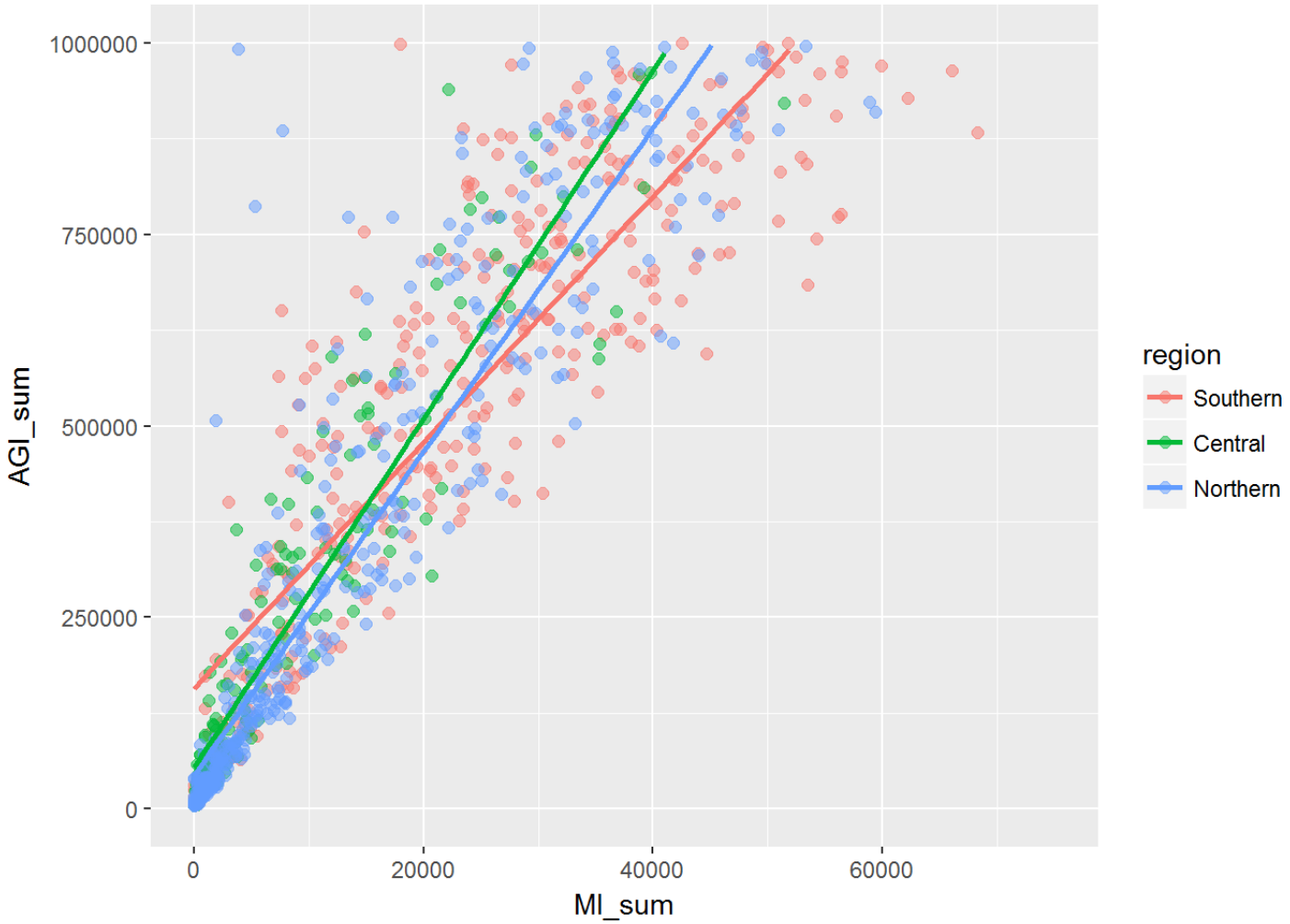
The mortage interest deduction and child and dependent care deduction were positively correlated with a correlation coefficient of 0.82. In addition, increasing ordinary dividends was correlated with increasing business income.

What was the strongest relationship you found?

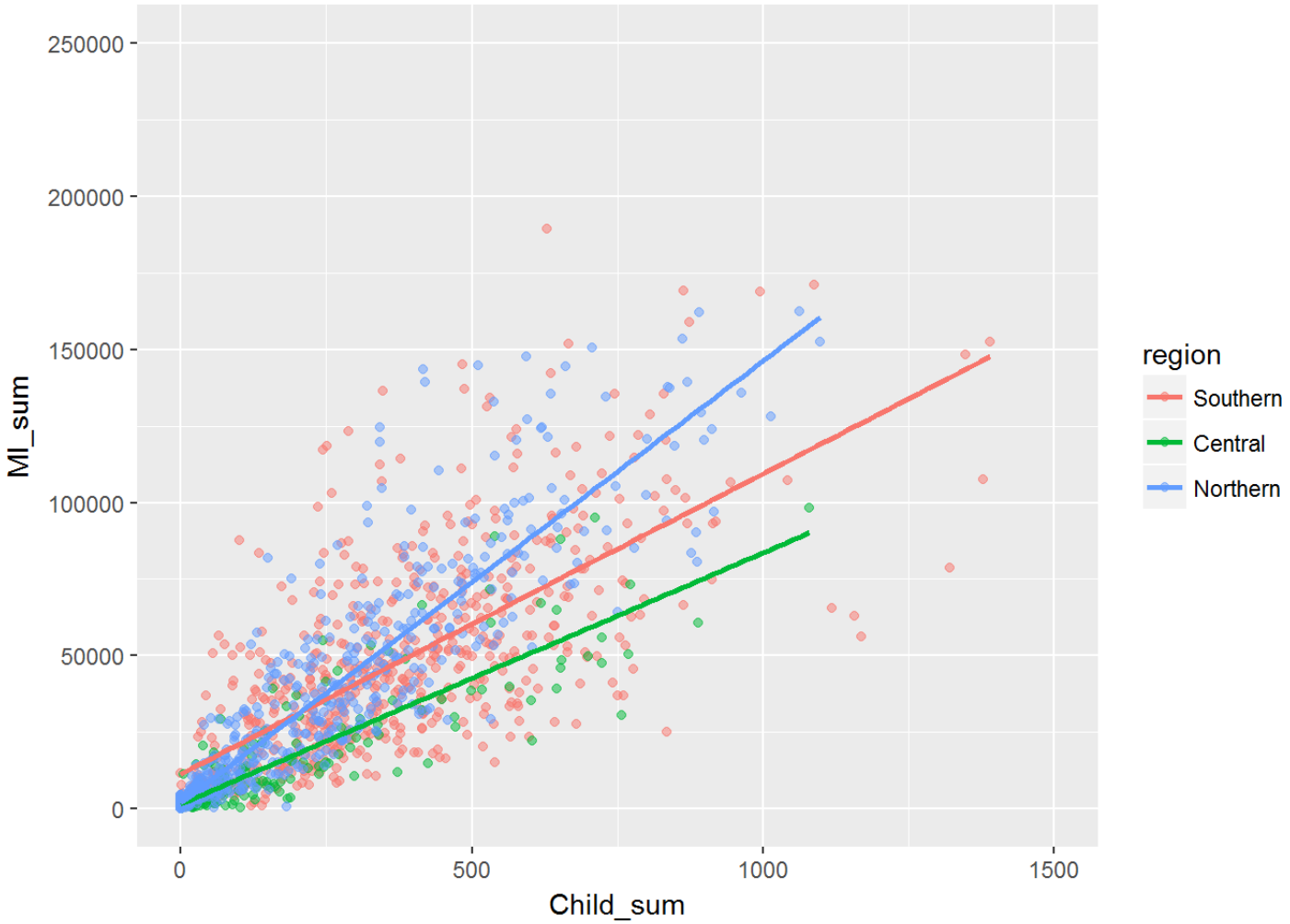
The AGI was strongly correlated with business income with a correlation coefficient of 0.91. Increases in business income for zipcode areas were correlated with an increase in AGI.

Multivariate Plots Section

This plot shows AGI_sum vs. Bus_inc_sum divided by region. Northern California zipcodes are concentrated in the < 25,000 Bus_inc_sum range while Southern California zipcodes are present in the full range of Bus_inc_sum values. Central California zipcodes are less represented in the dataset and are present in the full range of Bus_inc_sum.



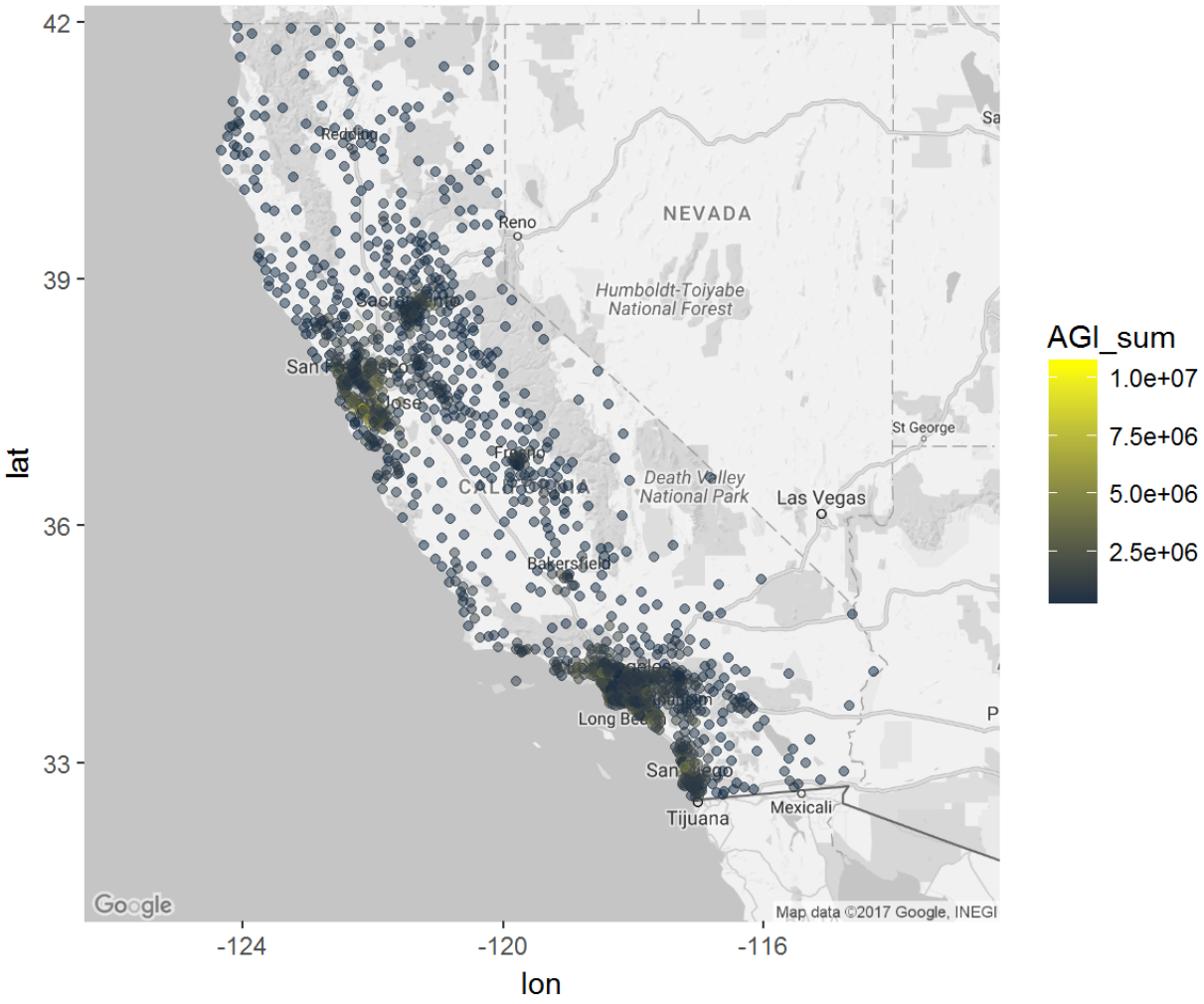
This plot of AGI_sum vs. MI_sum shows a similar pattern to the previous plot where Northern California zipcodes account for more of the MI_sum values less than 20,000 while Southern California values are present through the full range. Central California zipcodes are represented in the full range of MI_sum. The Southern California linear fit line is offset from the other regions.



This plot of MI_sum vs. Child_sum shows that all regions are represented in the range of Child_sum from 0 to 1000. There are differences in the linear fit models for the three regions.

This is a plot of Bus_inc_sum vs. Ord_div_sum divided by region. All regions are represented in the range of 0 to 50,000. The Southern California data is more offset from the Northern California data in the lower range of Ord_div_sum from 0 to 10,000. Central California data is sparsely represented in the dataset. The Southern California linear fit line is offset from the other two regions.

Next, the zipcode data was merged with latitude and longitude reference information to create a California map of AGI_sum variation.



The map plot above shows the AGI_sum by zipcode. Urban areas such as Los Angeles and San Francisco show high concentrations of zipcodes with greater AGI_sum.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

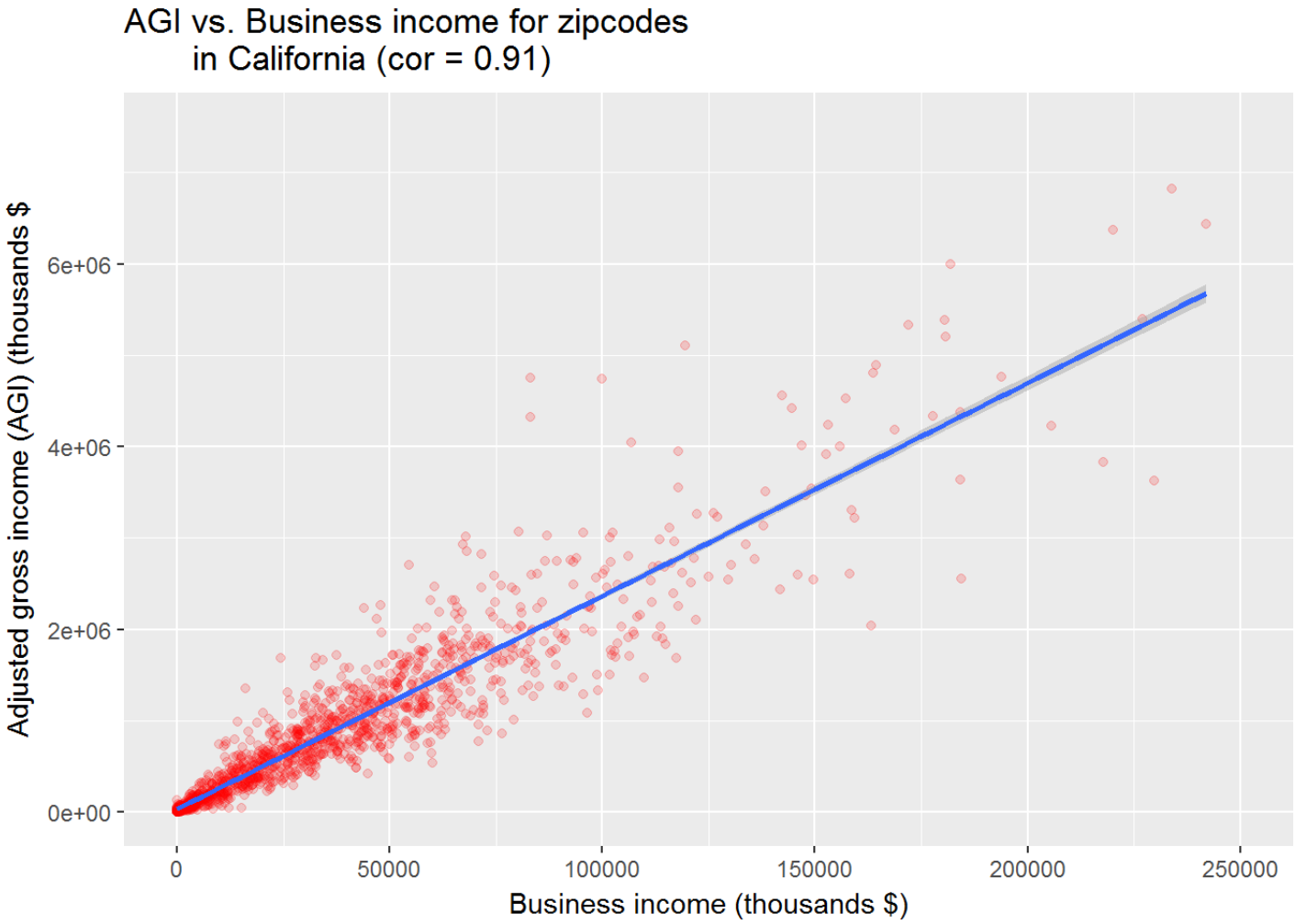
Northern California business income and mortgage interest deductions were concentrated in the lower range of these variables, while Southern California zipcodes were present in the full range. The distribution between regions was similar when plotting child care deduction vs. mortgage income. The positive correlation between business income and ordinary dividends was offset between Northern and Southern California zipcodes.

Were there any interesting or surprising interactions between features?

The map plot showed that higher AGI zipcodes were concentrated in urban and nearby suburban areas.

Final Plots and Summary

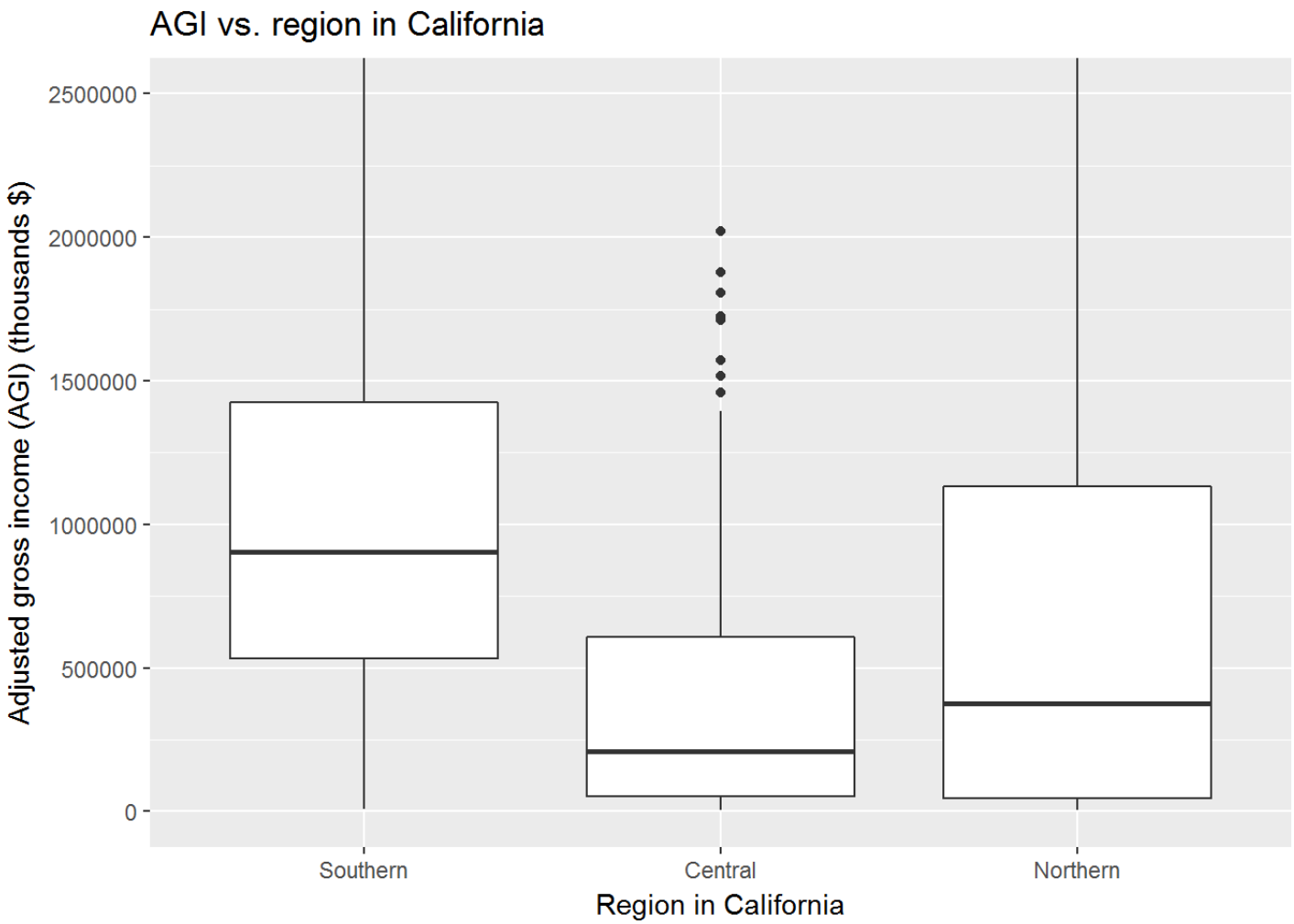
Plot One



Description One

This plot shows AGI vs. business income per zipcode in California with a linear model fit of the data. Outliers were excluded in this plot. The data is positively correlated as shown in the correlation coefficient of 0.91. As business income increases for a zipcode area, there is an associated increase in AGI.

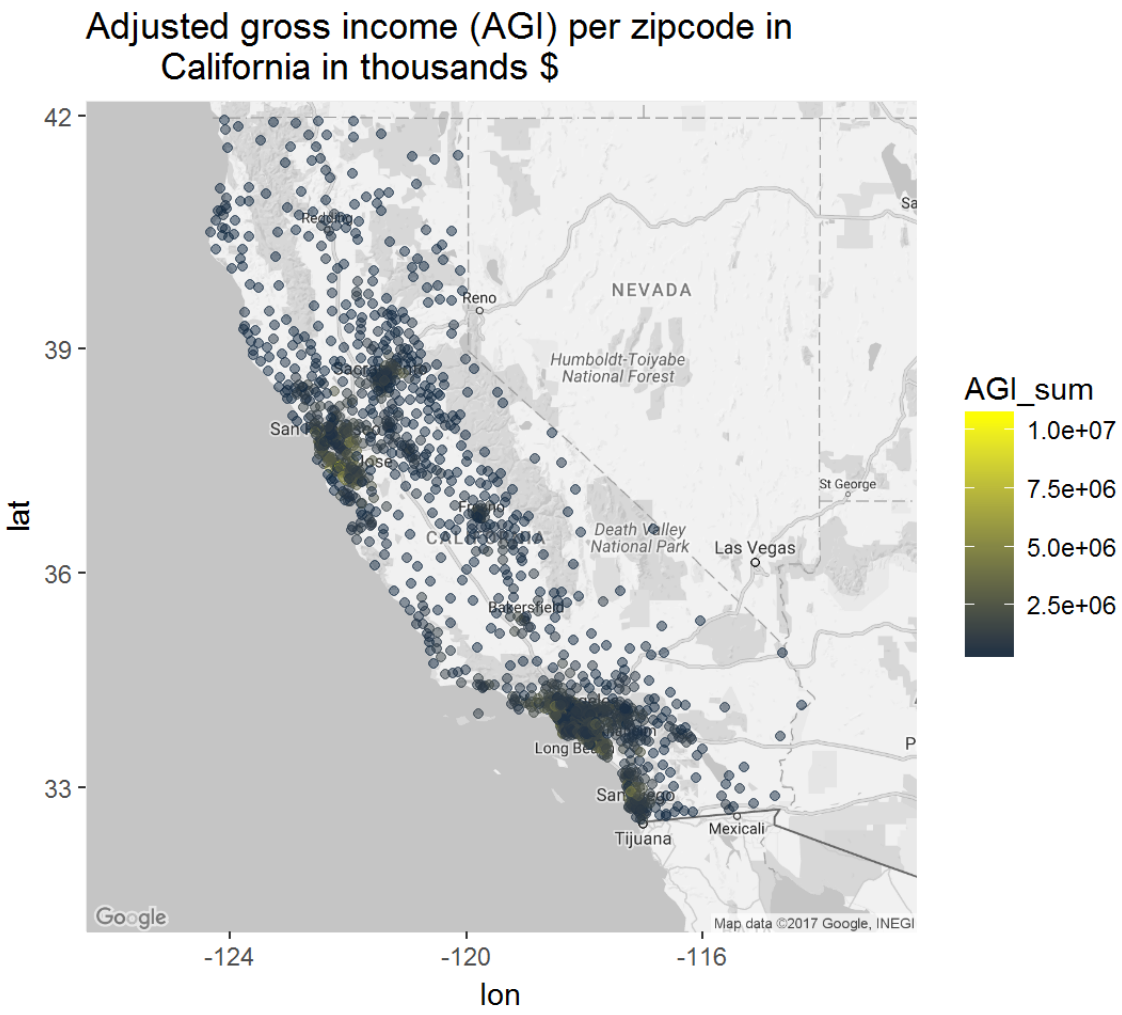
Plot Two



Description Two

The boxplot above shows the distribution of AGI for each region of California. Outliers were excluded in this plot. The median AGI is highest for Southern California (904,600) and lowest for Central California (374,400).

Plot Three



Description Three

The map plot above shows the AGI per zipcode in California. Metropolitan areas including Los Angeles, San Francisco, San Jose and Sacramento show greater concentrations of zipcodes with AGI > 7,500,000.

Reflection

The data was initially plotted to examine the distribution of variables by AGI category. However, dividing the zipcode data by AGI category would not provide an overview of each zipcode. This presented difficulties in understanding the data by zipcode. Therefore, a new dataframe was created to sum the data per zipcode and perform further analysis. This new dataframe was later successfully analyzed for data by zipcode. In order to create the map plot, a reference dataset was imported and merged with the existing dataset to add longitude and latitude data necessary for mapping.

Next, correlations were examined between variables in this new dataframe. Both business income and mortgage interest deduction showed positive correlation with AGI. A new variable was then created for the region of California that corresponds to each zipcode. Northern california values for mortage interest deduction, business income, ordinary dividends and child and dependent care deduction were more concentrated in the lower range of these variables, while Southern California values were present in the full

range of data. Central California zipcode values were less frequent in the dataset. Overall AGI was highest for Southern California and lowest for Central California. The map plot showed that urban areas contained more zipcodes with higher AGI.

This analysis could be expanded to account for population differences in zipcode areas and regions. After accounting for population differences, more accurate comparisons between variables could be made. In addition, this analysis could be performed after implementation of policies affecting home ownership, business development and investment as these were correlated with AGI in this dataset.