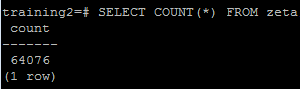
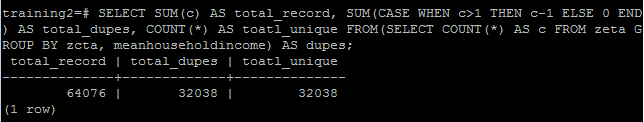
LAB 3: Analytic Methods

Retrieve and Clean Up Data

1. How many rows of data are there in the *zeta* table?

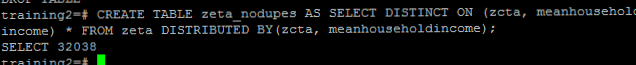


1. Are there any duplicate rows of data in the *zeta* table? If so, how can you tell?



Yes there are duplicates.

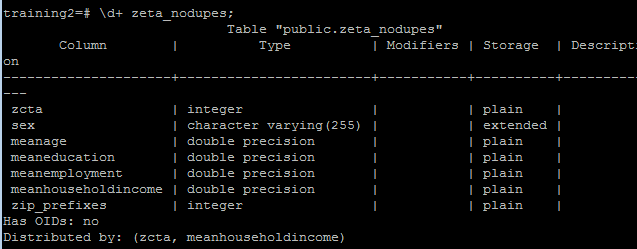
1. If there are duplicates, make a new table called *zeta\_nodupes* that has no duplicates. Now are there any duplicate rows of data? How can you tell?



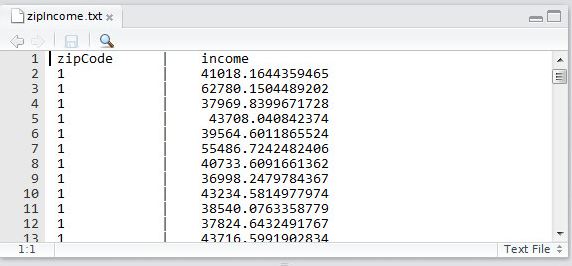
Group the data by the first character of the zip code

\*Create a new column in the *zeta\_nodupes* table called *zip\_prefixes* that contains the first digit of all of the zip codes in the *zcta* column for each row.



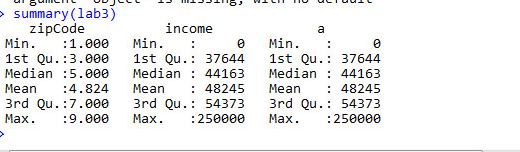


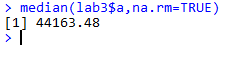
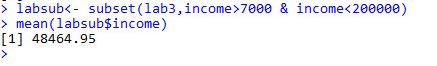
\*Take only the *zip\_prefix* and the *meanhouseholdincome* columns and save them into a separate file, ordered by zip code prefix, so that we can later analyze them in R. A filename like *zipIncome.txt* would be appropriate.



Data Analysis in R

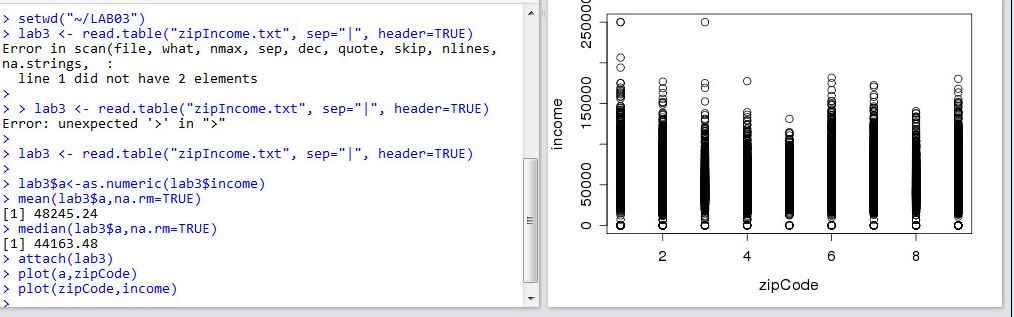
1. Analyze the summary of your data. What are the mean and median average incomes?





1. Plot a scatter plot of the data. Although this graph is not too informative, do you see any outlier values? If so, what are they?

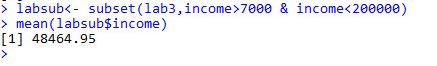
I don’t see any outlier values.



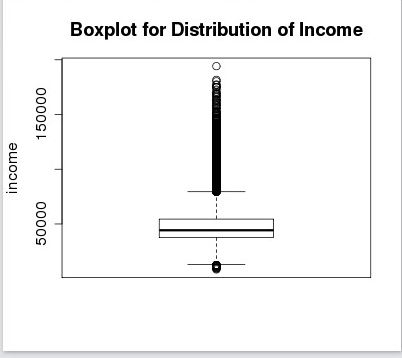
\* In order to omit outliers, create a subset of the data so that:

$7,000 < income < $200,000

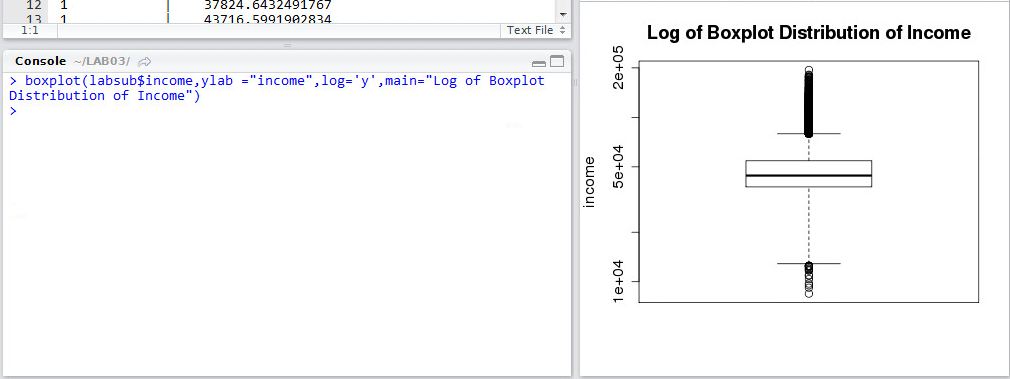
1. What’s your new mean?



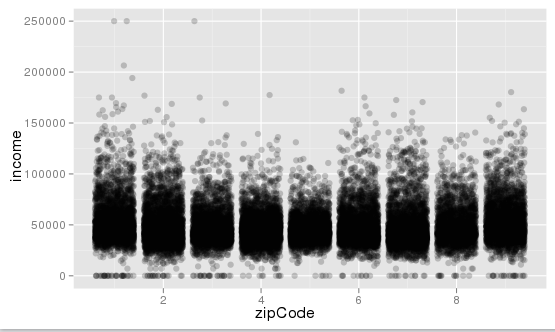
1. Create a simple box plot of your data. Be sure to add a title and label the axes.

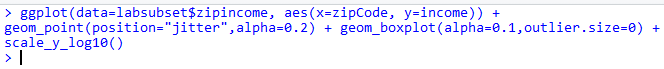


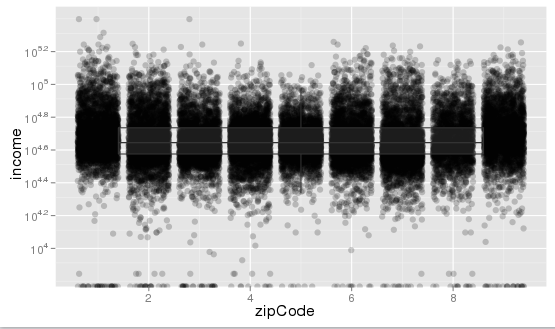
1. In the box plot you created, notice that all of the income data is pushed towards the bottom of the graph because most average incomes tend to be low. Create a new box plot where the y-axis uses a log scale. Be sure to add a title and label the axes.



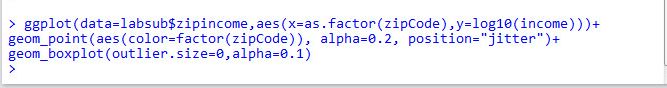
1. Make a *ggplot* that consists of just a scatter plot using the function *geom\_point()* with position = “*jitter”*  so that the data points are grouped by zip code. Be sure to use *ggplot*’s function for taking the log10 of the y-axis data. (Hint: for *geom\_point*, have *alpha*=0.2).

C:\Users\prash\AppData\Local\Microsoft\Windows\INetCache\Content.Word\jitter.png

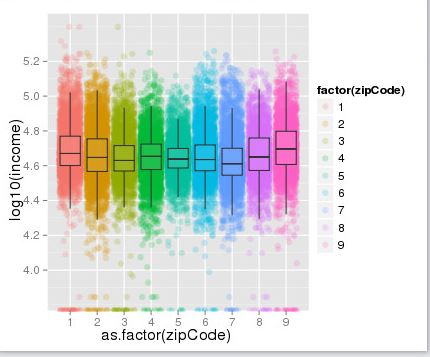




1. Create a new *ggplot* by adding a box plot layer to your previous graph. To do this, add the *ggplot* function *geom\_boxplot()*. Also, add color to the scatter plot so that data points between different zip codes are different colors. Be sure to label the axes and add a title to the graph. (Hint: for *geom\_boxplot*, have *alpha*=0.1 and *outlier.size*=0).



c



1. What can you conclude from this data analysis/visualization?

There were many duplicates which was removed and the grpah plotted depicted many practical things which could not be found out with the data. Data visualiztion provides the view where in we can see and understand how the data is and can be understood well.