Is There an Association Between Income Disparity and the Number of Bank Consumer Complaints?

Sriya Cheedella, Daisha Flowers, Hailin Yao

12/8/2019

Context

- When it comes to money, everyone has a love-hate relationship. People love it when they get their first paycheck, but once they begin to run out, they hate the concept of relying on a piece of paper to acquire their needs. Instead of looking within themselves, they turn to outside factors, such as their banks, to blame their problems on. However, banks are far from perfect. They make mistakes that need to be pointed out because they have a plethora of customers to handle. Due to this logic, banks receive numerous consumer complaints about various financial issues in hopes that the corresponding bank can solve them.

Problem

- We want to investigate if there is there is a certain cause behind customer complaints besides the issue stated with their bank. Out of all the probable causes, we wanted to examine if there is correlation between income disparity and the number of bank complaints. Since feelings are deeply intertwined with complaining, we wanted to utilize an objective variable that could aid us in discovering a variable that deeply influences the actions of bank customers.

Data Description

- Consumer Complaints Database
  - Collection of complaints about consumer financial products and services that CFBP sent to companies for response.
  - 1.7 million observations from 2012 to 2016.
  - Provided by Consumer Financial Protection Bureau (cfbp).
  - Each row is a complaint providing information about...
  - Columns: date, product, subproduct, issue, subissue, narrative, response, company, state, zipcode, tag, consent, submitted, datesent, privresponse, timely, disputed, complaintid and standard.

Data Description Cont.

- Average State Incomes
  - Provides state and its corresponding median income.
  - Collected by American Community Survey (ACS) in 2017, accounting for inflation.
  - Columns: GEO.id, GEO.id2, GEO.display-label, GRT_STUB.target-geo-id, GRT_STUB.target-geo-id2, GRT_STUB.rank-label, GRT_STUB.display-label, EST, and MOE.
- State Abbreviations
  - Provides state names and its corresponding abbreviation.
  - Collected by World Population Review.

Data Cleaning

- We used MySQL to remove the unneccessary columns because R wasn't able to load all the data.
- Data was very clean considering it was government data.
- Deleted rows that had zip code in wrong column (only a few rows).
- Zip code clean up:
  - Roughly half of the zip codes only had 2-3 digits (e.g. 203XX, 56XXX), so removed those rows.
- Deleted rows where the disputed column had "N/A".
- Began with 1.7 mil rows, left with 429126 rows at the end!

Data Cleaning with MySQL

- Commands used:
  - desc ccdata; -> Described the table ccdata, providing column name and its type.
  - select * from ccdata limit 100; -> Displayed the first 100 rows of the table ccdata.
  - alter table ccdata modify date varchar(255) not null; -> Changed the type of column "date" from table ccdata to character type and not null.
  - select distinct(product) from ccdata; -> Showed the distinct entries in the column "product" from table ccdata.
  - select count(*) from ccdata; -> Counted the total number of rows in table ccdata.
  - alter table ccdata drop narrative; -> Deleted the column "narrative" from table ccdata.
  - select * from ccdata where char_length(zipcode) = 5; -> Kept the rows with 5-digit zip codes.
  - select * from ccdata where disputed != "N/A"; -> Removed rows with N/A as the entry.

Data Cleaning with R

- With majority of the data cleaned in MySQL, we just read in the data and combine the useful information into a single dataframe:

```
#load complaints data
ccdata <- read.csv("/home/uscheed/Documents/temp/ccdata.csv", sep = ';',
                   header = FALSE)
colnames(ccdata) <- c("date", "product", "issue", "company", "state",
                      "zipcode", "response", "timely", "disputed",
                      "complaintid")
#load avg state incomes
avginc <- read.csv("/home/uscheed/Documents/temp/docs/avgStateIncomes.csv")
avginc <- avginc[c(7,8)]
index <- order(avginc$GRT_STUB.display.label)
avginc <- avginc[index, ]
avginc <- avginc[which(avginc$GRT_STUB.display.label != "Puerto Rico" &
                         avginc$GRT_STUB.display.label != "United States"),]
#load states and abbreviations
stateabb <- read.csv("/home/uscheed/Documents/temp/docs/state_abbrev.csv")
stateabb <- stateabb[c(1,3)]
#merge states and income to get abbreviations with avg income
finalavginc <- merge(stateabb, avginc, by.x = "State", by.y = "GRT_STUB.display
#final data frame
finaldf <- merge(ccdata, finalavginc, by.x = "state", by.y = "Code")
```

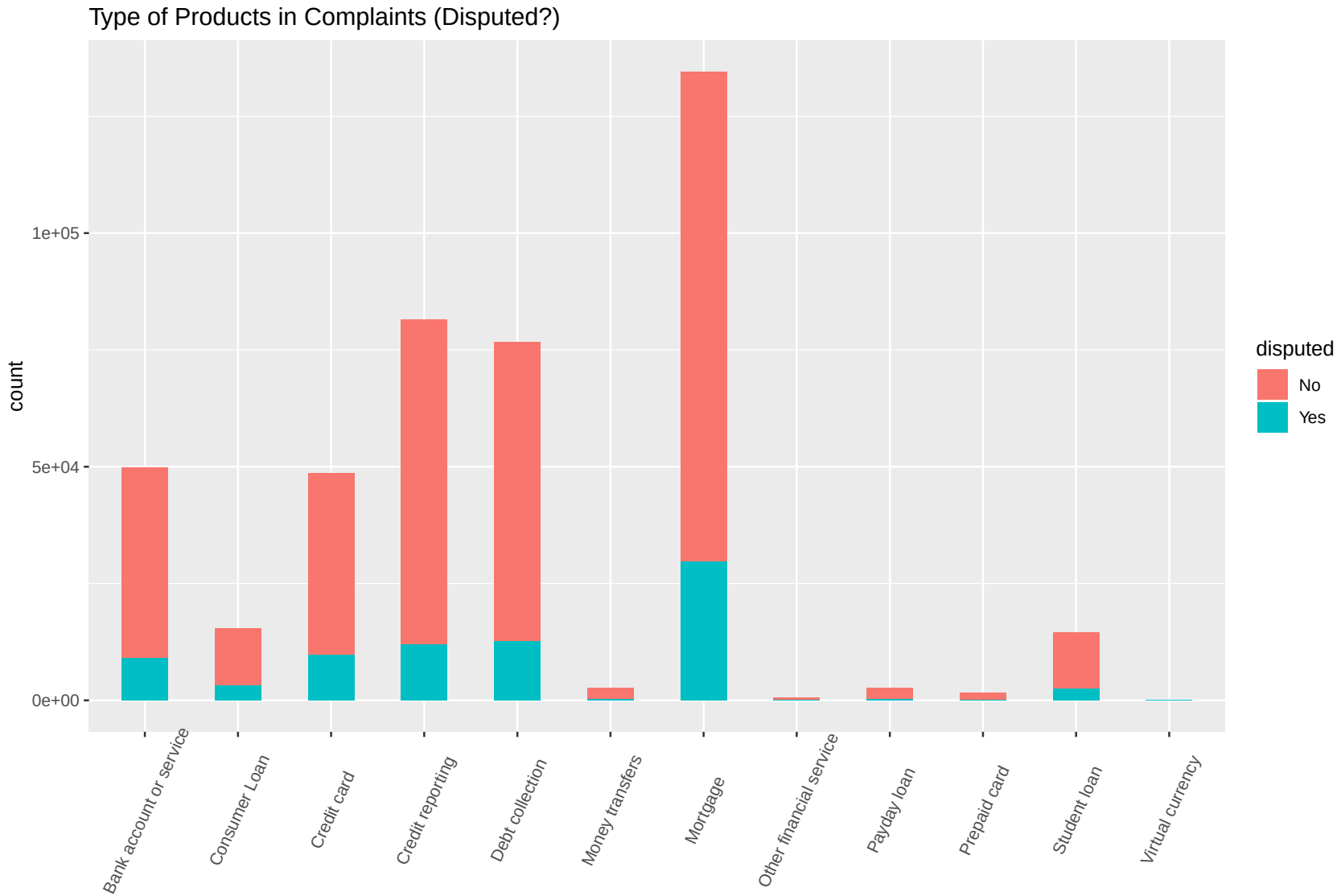
Final Dataframe

```
head(finaldf)
```

```
##   state       date         product                                    issue
## 1    AK 07/19/2014        Mortgage Loan modification,collection,foreclosure
## 2    AK 08/26/2014        Mortgage Loan modification,collection,foreclosure
## 3    AK 06/15/2012    Student loan      Problems when you are unable to pay
## 4    AK 12/28/2013 Debt collection          Disclosure verification of debt
## 5    AK 04/16/2016    Student loan       Dealing with my lender or servicer
## 6    AK 06/12/2015        Mortgage Loan servicing, payments, escrow account
##                                        company zipcode
## 1         BANK OF AMERICA, NATIONAL ASSOCIATION   99623
## 2         BANK OF AMERICA, NATIONAL ASSOCIATION   99504
## 3 ALASKA COMMISSION ON POST SECONDARY EDUCATION   99654
## 4       Receivables Performance Management, LLC   99503
## 5                         Navient Solutions, LLC.   99504
## 6                                    U.S. BANCORP   99654
##                          response timely disputed complaintid  State   EST
## 1         Closed with explanation    Yes       No      945031 Alaska 73181
## 2         Closed with explanation    Yes       No     1001227 Alaska 73181
## 3         Closed with explanation     No       No      102637 Alaska 73181
## 4         Closed with explanation    Yes      Yes      648163 Alaska 73181
## 5 Closed with non-monetary relief    Yes       No     1881822 Alaska 73181
## 6         Closed with explanation    Yes       No     1418765 Alaska 73181
```

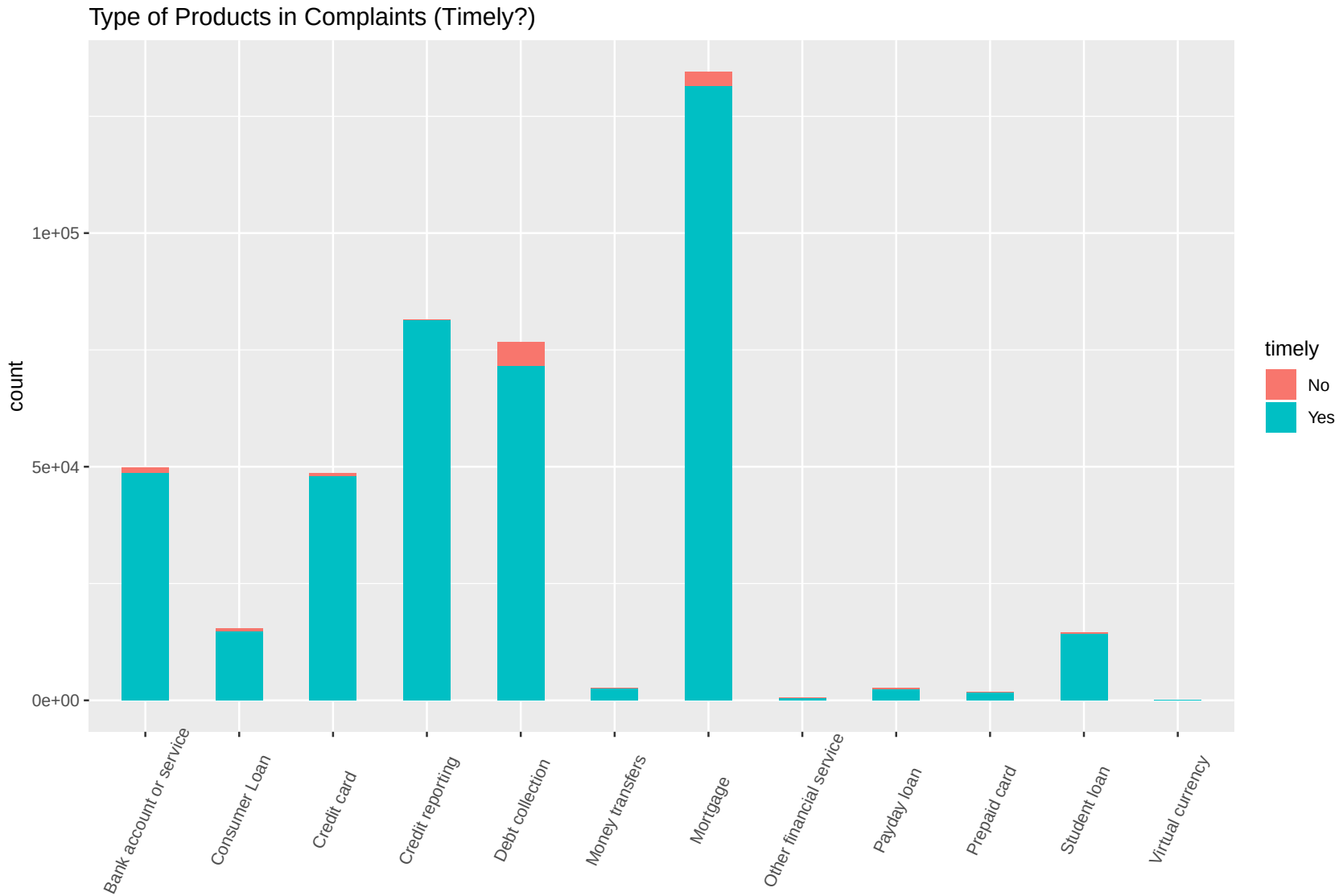
Exploratory Data Analysis

```
prod <- ggplot(finaldf, aes(product, fill = disputed))
prod + geom_bar(aes(x = product), width = 0.5) + theme(axis.text.x = element_text(angle=65, vjust=0.6)) + labs(title="Type of Products in Complaints (Disputed?
```



Exploratory Data Analysis Continued

```
time <- ggplot(finaldf, aes(product, fill = timely))
time + geom_bar(aes(x = product), width = 0.5) + theme(axis.text.x = element_text(angle=65, vjust=0.6)) + labs(title="Type of Products in Complaints (Timely?)"
```



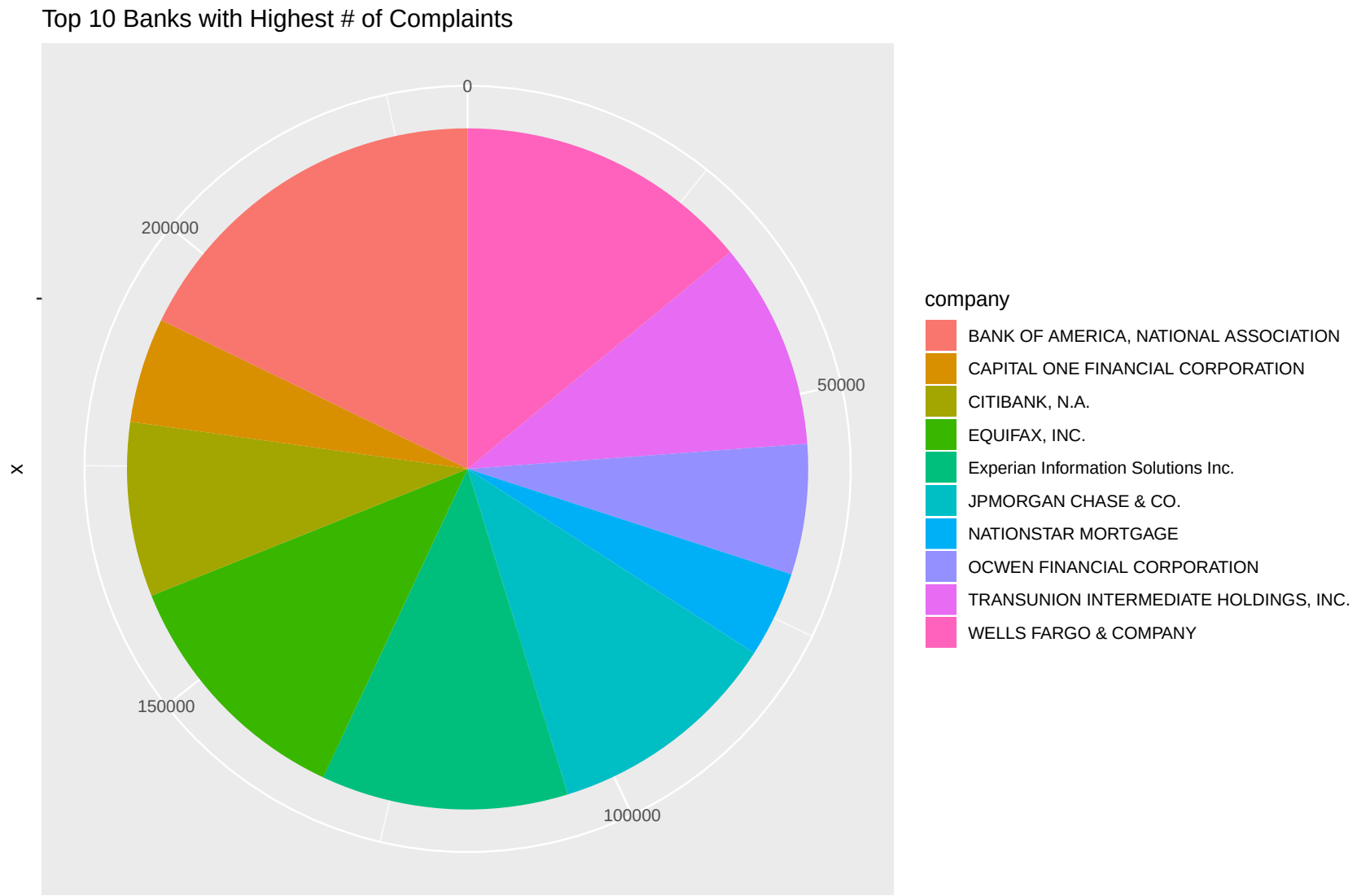
Exploratory Data Analysis Continued

```
bankcount <- count(finaldf, company)
bankind <- order(bankcount$n, decreasing = TRUE)
bankcount <- bankcount[bankind,]
bankcttop <- bankcount[1:10,]
bankcttop
```

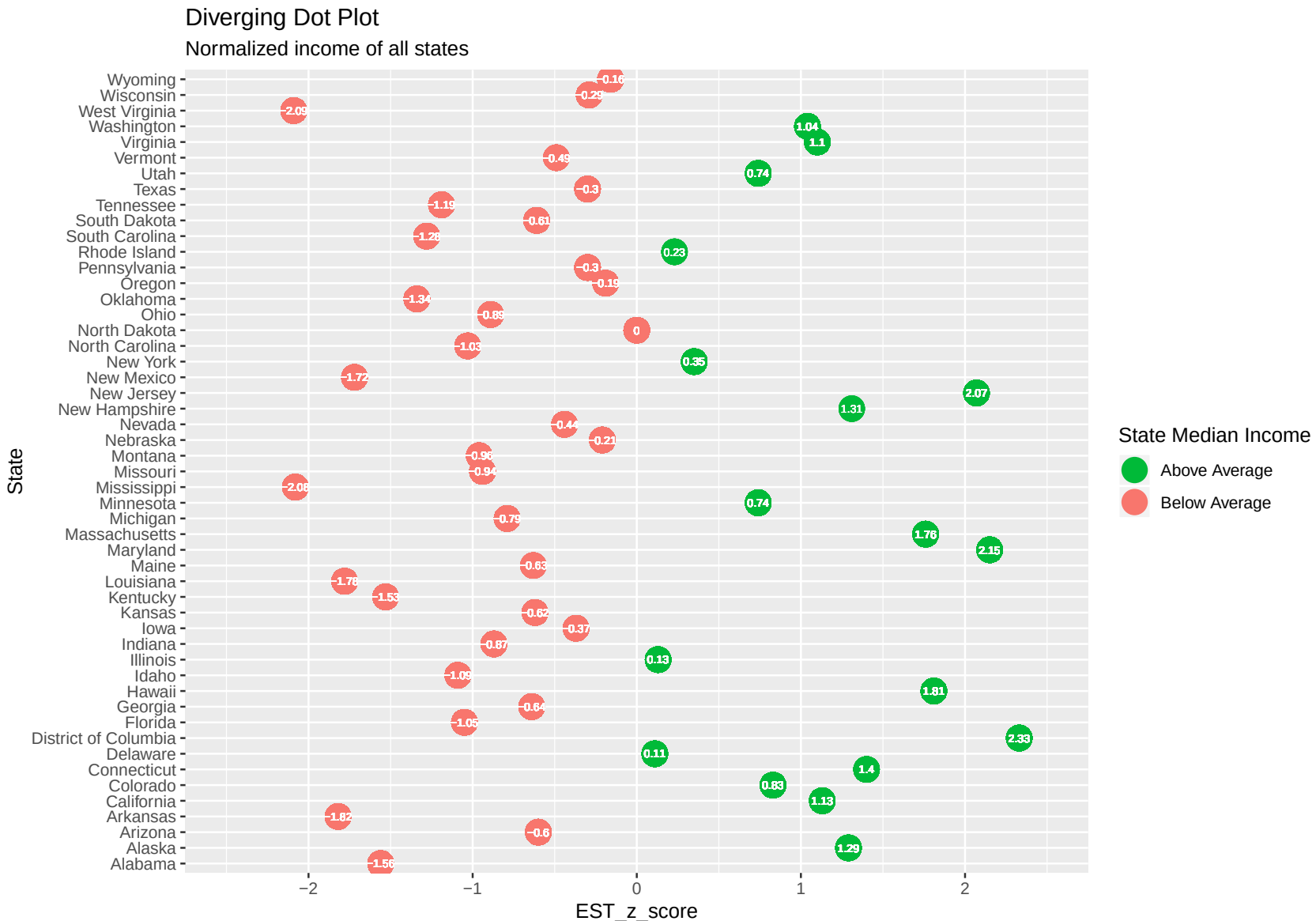
```
## # A tibble: 10 x 2
##    company                                    n
##    <fct>                                  <int>
##  1 BANK OF AMERICA, NATIONAL ASSOCIATION  41469
##  2 WELLS FARGO & COMPANY                  32651
##  3 EQUIFAX, INC.                          27844
##  4 Experian Information Solutions Inc.    27325
##  5 JPMORGAN CHASE & CO.                   25979
##  6 TRANSUNION INTERMEDIATE HOLDINGS, INC. 22836
##  7 CITIBANK, N.A.                         19299
##  8 OCWEN FINANCIAL CORPORATION            14413
##  9 CAPITAL ONE FINANCIAL CORPORATION      11601
## 10 NATIONSTAR MORTGAGE                     9503
```

Exploratory Data Analysis Continued

```
ggplot(bankcttop, aes(x = "", y=n, fill=company))+geom_bar(width = 1, stat="identity")+coord_polar("y", start = 0) +
  labs(title="Top 10 Banks with Highest # of Complaints")
```



Exploratory Data Analysis Continued



What is Simple Linear Regression?

- This method is used when we want to see if one independent variable is correlated with the dependent variable.
- Both variables are quantitative.
- Examples:
  - Are height and weight correlated?
  - Is money spend on advertising associated with the company's profit?

SLR Assumptions

- The data is normally distributed.
- We have over 30 observations in our dataset, so by the Central Limit Theorem our data is normally distributed.
- There is a linear relationship between the independent and dependent variables.

```
compcts <- count(finaldf, vars=State)
compcts$vars <- tolower(compcts$vars)
ctsinc <- merge(compcts, finalavginc, by.x = "vars", by.y = "State")
ctsinc <- ctsinc[, c(1,2,4)]
#ggplot(data = ctsinc, aes(x=n, y=EST)) + geom_point() + labs(title = "Income vs. # of Comp
# cor(ctsinc$EST, ctsinc$n)
```

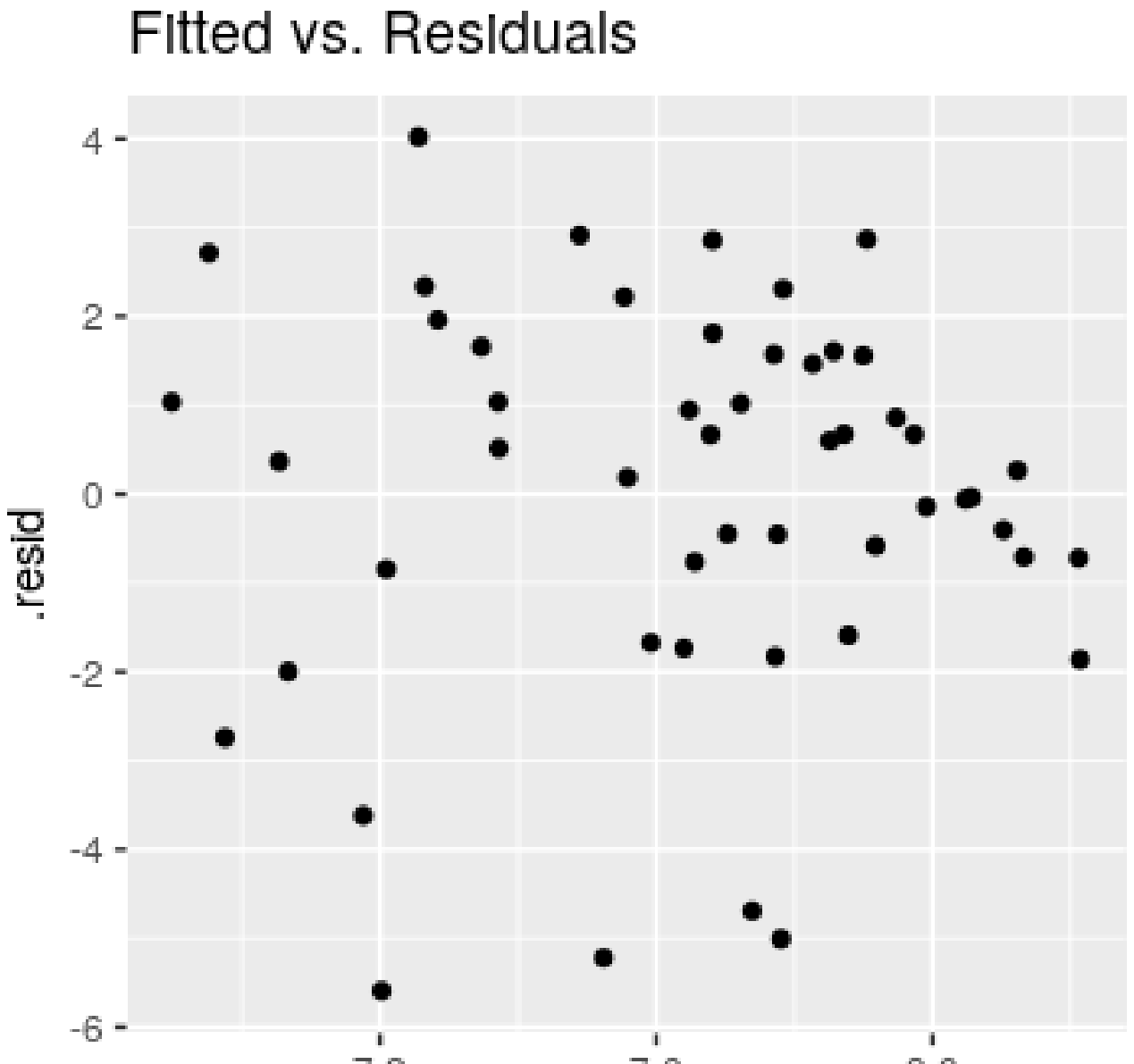
0.09541067

```
# cor(ctsinc$EST, ctsinc$n_log)
```

-0.156537

- There is barely a linear relationship even after the log transformation, but we will continue anyways.
- The variability of the residuals will be relatively constant across all values of X.

SLR Assumptions Continued



SLR Assumptions Continued

- There seems to constant variance (homoscedasticity) so we can continue.
- The y-values are independent of each other.
- We can assume that the number of complaints from each state are independent of each other since the data collection doesn't rely on state when the complaints are being recorded. The randomness of the residual plot verifies this reasoning.

Simple Linear Regression

```
# cclm <- lm(n_log ~ EST, data = ctsinc)
# summary(cclm)
```

Call: lm(formula = n_log ~ EST, data = ctsinc)

Residuals: Min 1Q Median 3Q Max -5.5927 -0.8041 0.3673 1.5667 4.0197

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.684e+00 1.862e+00 5.201 3.87e-06 *** EST -3.382e-05 3.048e-05 -1.109 0.273
— Signif. codes: 0 ‘’ ***0.001*** ‘’ *0.01* ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.21 on 49 degrees of freedom Multiple R-squared: 0.0245, Adjusted R-squared: 0.004597
F-statistic: 1.231 on 1 and 49 DF, p-value: 0.2726

- The equation produced is:
- $(NumComplaints) = -0.00003382(Income) + 9.684$
- With an R^2 of 0.0245, we can see that the model fits the data horribly. Plus, we can see that the state's median income is not a significant predictor of the number of complaints because its p-value is greater than 0.05.
- There is sufficient evidence to conclude that there is no association between a state's number of complaints and its median income.

What is Multinomial Logistic Regression?

- This supervised machine learning algorithm is used when we want to find an association between predictor(s) (x) and three or more categorical explanatory (y) variables.
- The response variables are unordered, meaning that one variable is not more important than the others (there is no natural order).
- Examples:
  - Is there an association between student's grades in high school and which major they choose in college?
    - X: Percentage grades in history, math, science, english.
    - Y: History, Math, Science, English.
  - Is someone's mood related to their daily activities?
    - X: Time spent in social activities, time spent in exercise, time spent in studying, etc.
    - Y: Happiness, sadness, anger, fear, disgust, surprise.

MLR Assumptions

- For multinomial logistic regression, we do not have to check for normality, linearity or homoscedasticity.
- We do have to check for multicollinearity, but we only have one predictor.
  - Multicollinearity is when the independent variables in a model are linearly correlated with each other.
  - For example, race length and race time & age and diseases.

Multinomial Logistic Regression

```
finaldf$product2 <- relevel(finaldf$product, ref = "Mortgage")
ccmod <- multinom(product ~ EST, data=train)
```

```
## Warning in multinom(product ~ EST, data = train): group 'Virtual currency' is
## empty
```

```
## # weights:  33 (20 variable)
## initial  value 719517.251350
## iter  10 value 565037.080394
## iter  20 value 547553.684413
## iter  30 value 542819.865238
## final  value 542818.319010
## converged
```

```
summary(ccmod)
```

```
## Call:
## multinom(formula = product ~ EST, data = train)
##
## Coefficients:
##                          (Intercept)           EST
## Consumer Loan            -0.34032900 -1.354518e-05
## Credit card              -0.04855471  4.810105e-07
## Credit reporting          1.19092328 -1.134538e-05
## Debt collection           1.11066662 -1.098771e-05
## Money transfers          -3.13691288  3.105571e-06
## Mortgage                  0.72205710  4.387201e-06
## Other financial service -3.93430251 -1.015301e-05
## Payday loan              -1.53072686 -2.278747e-05
## Prepaid card             -3.21890417 -2.984702e-06
## Student loan             -0.76181107 -7.563378e-06
##
## Std. Errors:
##                           (Intercept)           EST
## Consumer Loan            2.853669e-12 1.780670e-07
## Credit card              1.902154e-12 1.212585e-07
## Credit reporting         1.725358e-12 1.090191e-07
## Debt collection          1.745522e-12 1.102541e-07
## Money transfers          5.974248e-12 3.802068e-07
## Mortgage                 1.535094e-12 9.976338e-08
## Other financial service 1.358142e-11 8.500834e-07
## Payday loan              6.311223e-12 3.889438e-07
```

Our Model's Accuracy

```
pR2(ccmod)
```

```
## fitting null model for pseudo-r2

## Warning in multinom(formula = product ~ 1, data = train): group 'Virtual
## currency' is empty

## # weights:  22 (10 variable)
## initial  value 719517.251350
## iter  10 value 573052.531185
## iter  20 value 543475.560635
## iter  30 value 543415.969465
## final  value 543410.484065
## converged

##           llh       llhNull            G2      McFadden          r2ML
## -5.428183e+05 -5.434105e+05  1.184330e+03  1.089720e-03  3.939172e-03
##          r2CU
##  4.047356e-03
```

```
wald <- summary(ccmod)$coefficients/summary(ccmod)$standard.errors
p_ccmod <- (1 - pnorm(abs(wald), 0, 1)) * 2
p_ccmod
```

```
##                         (Intercept)          EST
## Consumer Loan                     0 0.000000e+00
## Credit card                       0 7.283887e-05
## Credit reporting                  0 0.000000e+00
## Debt collection                   0 0.000000e+00
## Money transfers                   0 2.220446e-16
## Mortgage                          0 0.000000e+00
## Other financial service           0 0.000000e+00
## Payday loan                       0 0.000000e+00
## Prepaid card                      0 4.150829e-10
## Student loan                      0 0.000000e+00
```

- With a McFadden's R^2 of 0.00105, we see that the model fits the data horribly. However, we notice that all the product types are less than 0.05 except prepaid card, which means that the average state income is a significant predictor for those products.
- There is sufficient evidence that there is no association between the type of bank consumer complaints and the state income

Drawbacks

- We use the median state income, which encompasses the income disparities among all counties. Counties in certain areas of a state have a higher median income than others. Prime example is Virginia; northern Virginia counties have much higher household incomes compared to southern Virginia. Consequently, the median state income may differ greatly from their actual income and utilizing the county's income is a more accurate representation.
- While cleaning the data, we removed the zip codes that aren't five digits long. However, when R read the data, it removes the leading zeros, making it an invalid zip code (00568 -> 568). This may have removed vital data and made the distribution more skewed and less normally distributed.
- We are not accounting for the population of each state. This might be an issue because if there are more people in a state, this leads to a higher number of complaints.
- Complaints depend on the personalities of the people too, and that could be dependent on the state's culture and the upbringing of the person itself. A person can have a low income but be satisified with their financial state, while an upper middle class person can be stingy.

Number of Complaints Per State (Top 10)

```
#long and lat of states
states <- map_data("state")
states <- states[c(1,2,3,5)]
finalavginc$State <- tolower(finalavginc$State)
mapinc <- merge(finalavginc, states, by.x = "State", by.y = "region")

ziplonglat <- read.csv("~/Documents/temp/docs/zipcodelonglat.txt")
mapzip <- merge(finaldf, ziplonglat, by.x = "zipcode", by.y = "ZIP")
mapzip <- mapzip[which(mapzip$state != "ak" & mapzip$state != "hi" & mapzip$LNG > -130),]
```

```
compcounts <- count(mapzip, vars=State)
ccindex <- order(compcounts$n, decreasing = TRUE)
compcounts <- compcounts[ccindex,]
compcounts
```

```
## # A tibble: 51 x 2
##    vars             n
##    <fct>        <int>
##  1 California   74412
##  2 Florida      44785
##  3 Texas        35738
##  4 New York     32285
##  5 Georgia      23099
##  6 Illinois     17220
##  7 Maryland     15280
##  8 Virginia     14196
##  9 Pennsylvania 13052
## 10 Ohio         12729
## # ... with 41 more rows
```

Final Visualization - US Map