Income
in the United States

Jake Lieberfarb, Luis Ahumada,
Noah Feldman, Becca Blacker

Contents

Question

How predictable is income in American communities?

How can we predict it?

Income per capita

Dataset

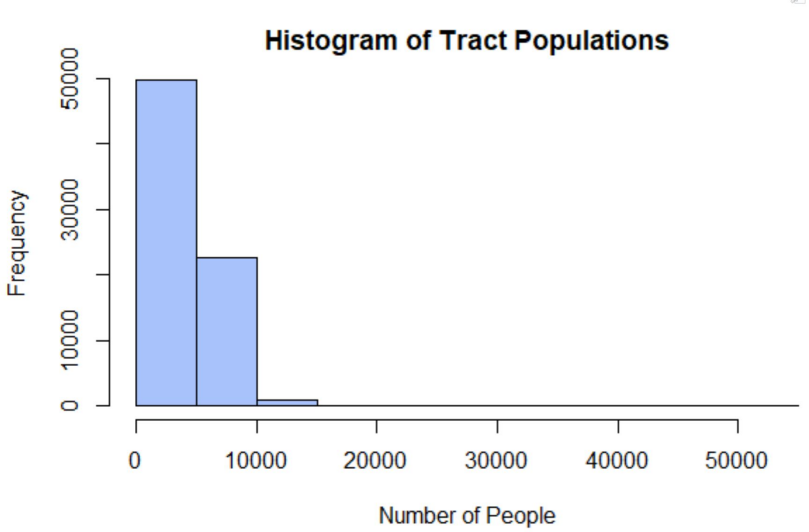
Census Tracts

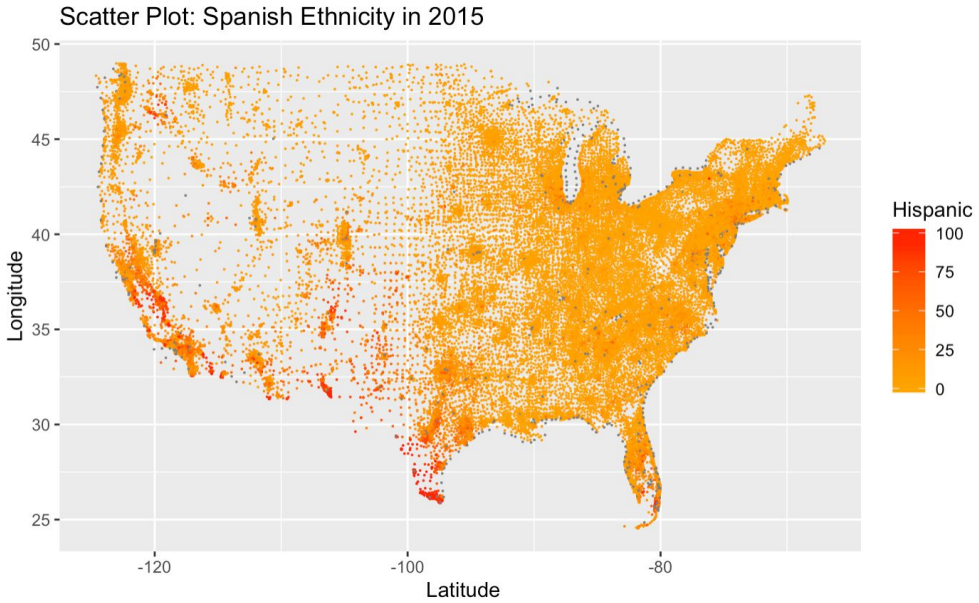
- 2015
- 37 variables
- 74,001 instances
- American Community Survey 5-year estimates

United States Census Bureau

kaggle

Census Tract Distribution





Variables

Independent

Work Variation

- Professional
- Service
- Office
- Construction
- Production
- Unemployed
- Self Employed

Sums to 1

Ethnic Variation

- White
- Black
- Hispanic
- Asian

Sums to ~ 1

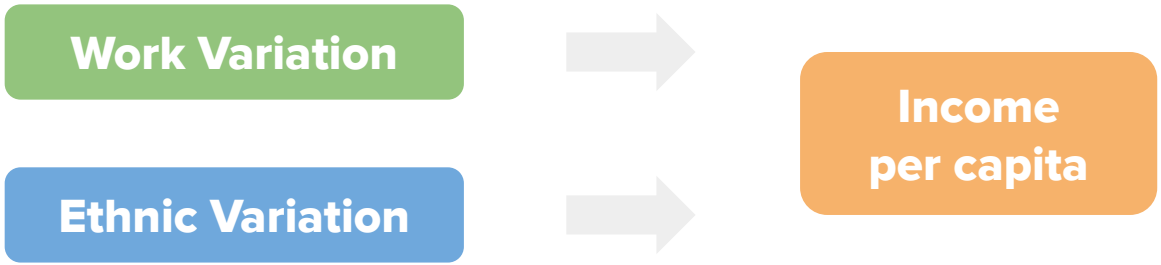
Dependent

Income

- Income Per Capita
- Low/High Income
- Low/Mid-Low/Mid-High/High Income

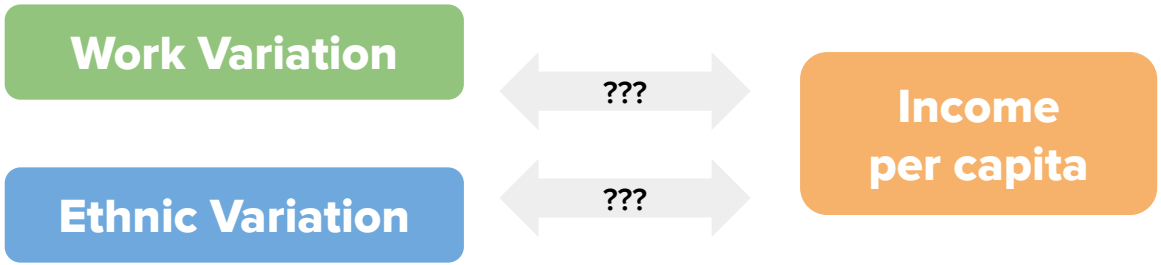
Question

How do work and ethnicity in a community predict income?



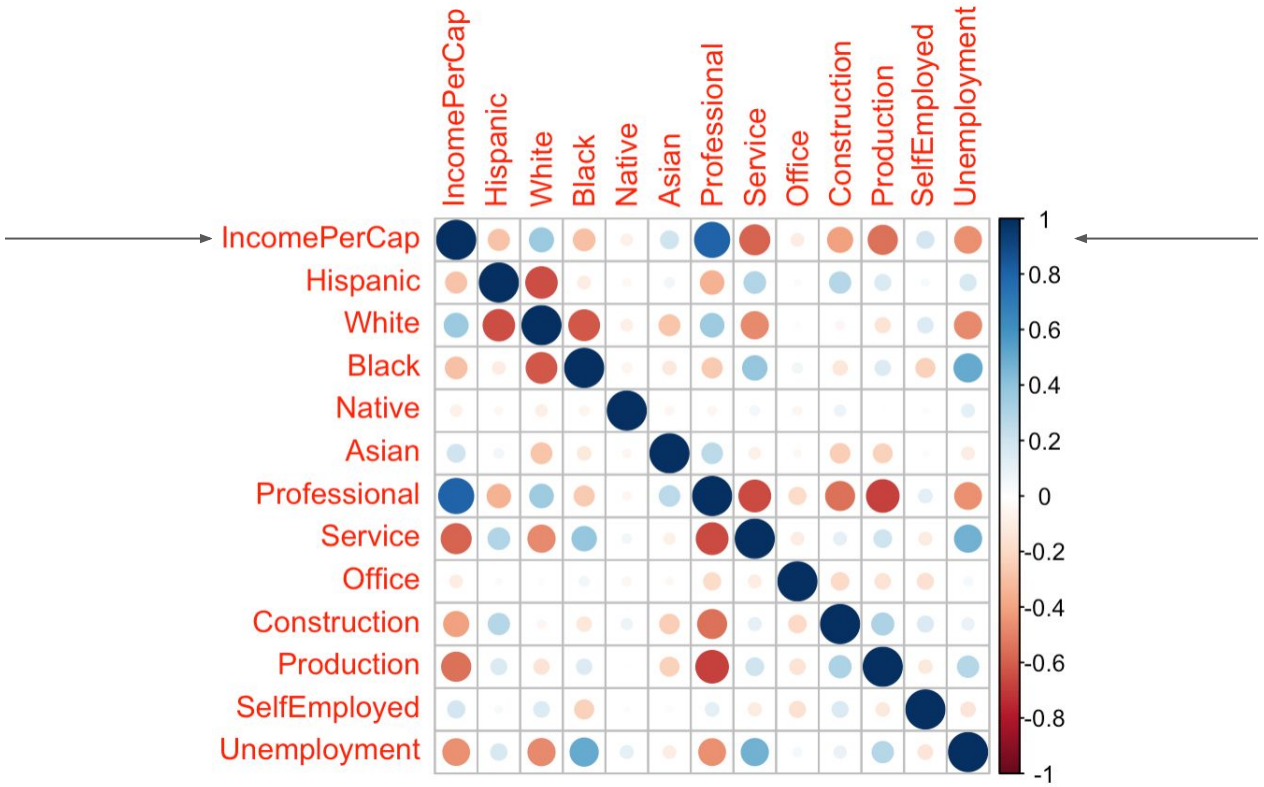
Causality

- Data forms a snapshot in 2015
- Hard to determine directionality



- Data exists for other years, a future, more temporal analysis could show causality.

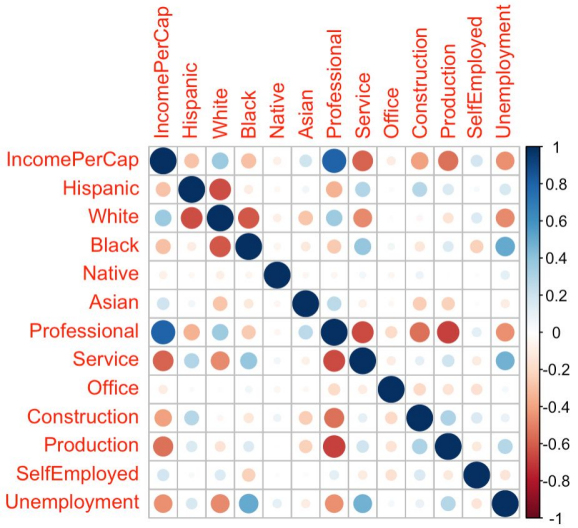
Correlation Matrix



Principal Component Analysis/Regression

Corr/ Scaled Cov Matrix

- Goal of PCA:
  - To decrease the number of variables while accounting for the variation of the data.
  - Account for collinearity in the predictors.



	Hispanic	White	Black	Native	Asian	Professional	Service	Office	Construction	Production	SelfEmployed	Unemployment
Hispanic	1.000	-0.641	-0.097	-0.036	0.058	-0.350	0.296	-0.022	0.285	0.158	0.031	0.165
White	-0.641	1.000	-0.611	-0.081	-0.274	0.354	-0.474	-0.013	-0.041	-0.140	0.150	-0.473
Black	-0.097	-0.611	1.000	-0.053	-0.111	-0.255	0.380	0.059	-0.126	0.143	-0.228	0.509
Native	-0.036	-0.081	-0.053	1.000	-0.043	-0.044	0.052	-0.044	0.074	0.005	0.011	0.102
Asian	0.058	-0.274	-0.111	-0.043	1.000	0.267	-0.077	-0.032	-0.240	-0.224	-0.018	-0.096
Professional	-0.350	0.354	-0.255	-0.044	0.267	1.000	-0.652	-0.184	-0.541	-0.687	0.112	-0.458
Service	0.296	-0.474	0.380	0.052	-0.077	-0.652	1.000	-0.097	0.100	0.196	-0.103	0.471
Office	-0.022	-0.013	0.059	-0.044	-0.032	-0.184	-0.097	1.000	-0.192	-0.148	-0.160	0.043
Construction	0.285	-0.041	-0.126	0.074	-0.240	-0.541	0.100	-0.192	1.000	0.309	0.156	0.090
Production	0.158	-0.140	0.143	0.005	-0.224	-0.687	0.196	-0.148	0.309	1.000	-0.111	0.286
SelfEmployed	0.031	0.150	-0.228	0.011	-0.018	0.112	-0.103	-0.160	0.156	-0.111	1.000	-0.130
Unemployment	0.165	-0.473	0.509	0.102	-0.096	-0.458	0.471	0.043	0.090	0.286	-0.130	1.000

PCA

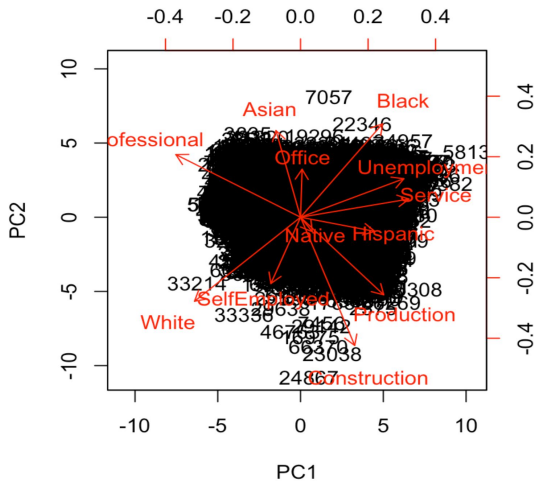
Importance of components:

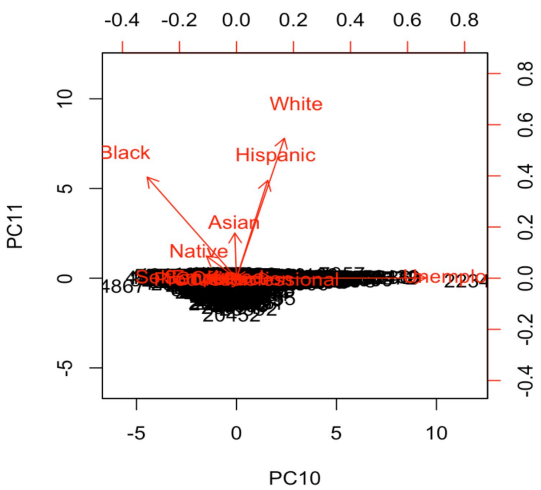
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	1.8450	1.3526	1.2005	1.05028	0.99900	0.94356	0.82517	0.79757	0.73865	0.68328	0.06624	0.003138
Proportion of Variance	0.2837	0.1525	0.1201	0.09192	0.08317	0.07419	0.05674	0.05301	0.04547	0.03891	0.00037	0.000000
Cumulative Proportion	0.2837	0.4361	0.5562	0.64815	0.73132	0.80551	0.86225	0.91526	0.96073	0.99963	1.00000	1.000000

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Hispanic	0.274996718	-0.05841443	0.55704472	0.32731074	-0.07592522	0.020222212	-0.34948546	0.14737207	-0.32634667	0.137020566	0.4777284029	-1.383495e-04
White	-0.392529745	-0.34695993	-0.34537009	0.00525174	0.03313168	-0.016102150	0.06082807	-0.29786235	0.02788939	0.209972588	0.6838024340	-1.810751e-04
Black	0.303734433	0.38585984	-0.27707022	-0.24413914	0.20566425	0.181113753	-0.01475000	0.35051362	0.16509224	-0.391399349	0.4938213300	-1.246822e-04
Native	0.043526994	-0.05956010	0.02210707	-0.47231661	-0.83613175	-0.177630246	0.01254622	0.04420382	-0.09526804	-0.131835582	0.1083950606	-3.546542e-05
Asian	-0.091018464	0.35793329	0.47585457	0.05204827	0.01757812	-0.322099977	0.60443506	-0.17725243	0.29163659	-0.007463110	0.2202899170	-6.296813e-05
Professional	-0.461559381	0.25885194	0.12542434	-0.21333169	0.07207381	0.005486454	-0.19842895	0.22649798	-0.07233603	0.159404405	-0.0014795540	7.295014e-01
Service	0.401381673	0.07523378	0.03041217	-0.13937797	0.04300042	0.136858861	-0.15945008	-0.76731153	0.01307506	-0.131243879	0.0059713171	4.008482e-01
Office	0.005765095	0.19811238	-0.27998739	0.66150226	-0.45020711	0.320813624	0.22749637	0.03631192	0.02674205	-0.065766062	0.0014641743	2.846342e-01
Construction	0.200622136	-0.52973946	0.14695070	0.03942497	-0.04015811	0.003124123	-0.05130285	0.22744611	0.71685092	-0.007914188	-0.0007771851	2.933486e-01
Production	0.308113225	-0.32058256	-0.17925722	0.03112861	0.18730559	-0.403728529	0.42471952	0.17443863	-0.45521843	-0.113940087	-0.0032966338	3.742703e-01
SelfEmployed	-0.109800919	-0.27379026	0.31198686	-0.24418300	0.05833854	0.722269880	0.41308892	0.03722832	-0.20567696	-0.128888110	-0.0008820205	5.795922e-06
Unemployment	0.382683747	0.15932342	-0.14357544	-0.20681792	-0.03981243	0.157691464	0.19552254	0.09161825	0.04261489	0.830596225	0.0009604257	1.466904e-05

PC11	PC12
Min. :-2.02621	Min. :-9.920e-03
1st Qu.:-0.01925	1st Qu.:-3.964e-05
Median : 0.01357	Median :-8.306e-06
Mean : 0.00000	Mean : 0.000e+00
3rd Qu.: 0.03477	3rd Qu.: 3.026e-05
Max. : 0.25752	Max. : 1.006e-02

PCA/Biplots



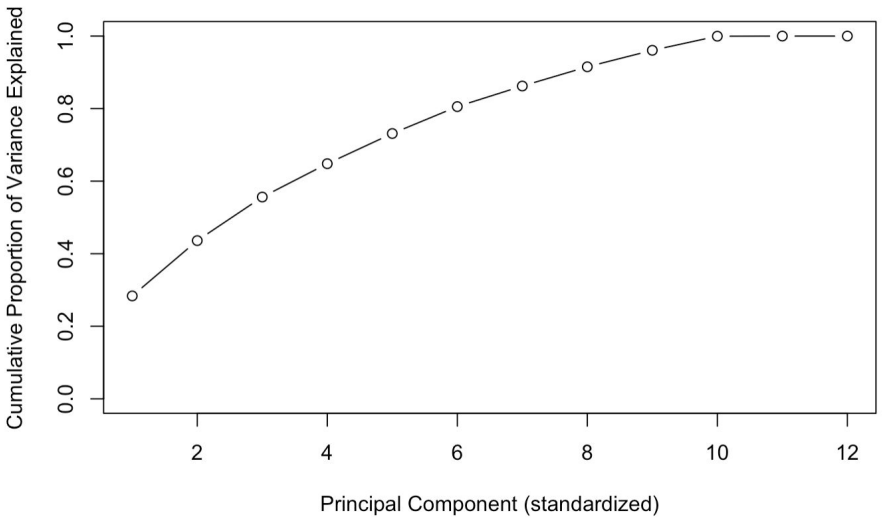


PC1		PC2	
Min.	:-6.2474	Min.	:-10.804786
1st Qu.	:-1.2715	1st Qu.	: -0.939785
Median	:-0.2526	Median	: -0.009147
Mean	: 0.0000	Mean	: 0.000000
3rd Qu.	: 1.1051	3rd Qu.	: 0.848478
Max.	:10.3909	Max.	: 8.111678

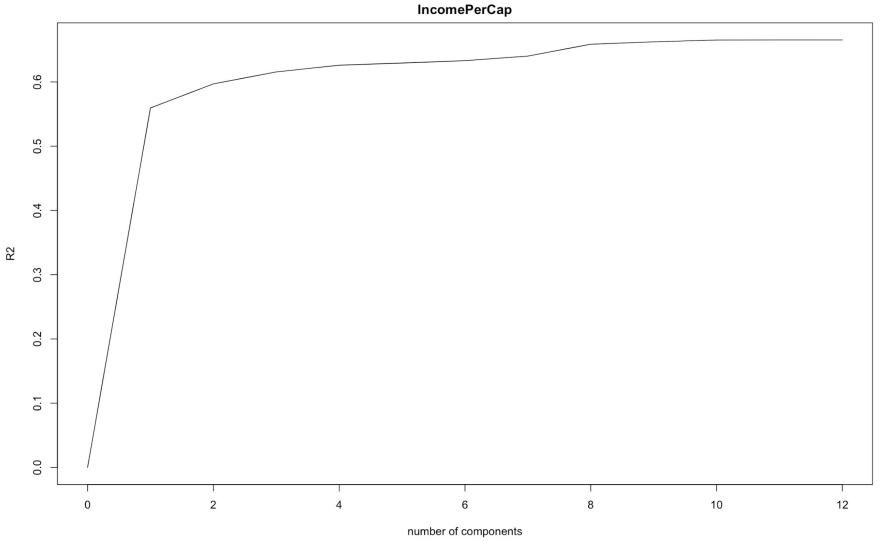
PC10		PC11	
Min.	:-5.99388	Min.	:-2.02621
1st Qu.	:-0.38812	1st Qu.	:-0.01925
Median	:-0.01357	Median	: 0.01357
Mean	: 0.00000	Mean	: 0.00000
3rd Qu.	: 0.36695	3rd Qu.	: 0.03477
Max.	:11.84471	Max.	: 0.25752

PCA

Cumulative Proportion of Variance



R Squared



PCA

```
Call:
lm(formula = IncomePerCap ~ ., data = comp_pcr_rot)

Residuals:
    Min      1Q  Median      3Q     Max
-103963   -4374    -459    3390  193391

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28688.2       32.1  892.73  < 2e-16 ***
PC1          -6049.0       17.4 -347.29  < 2e-16 ***
PC2           2134.8       23.8   89.86  < 2e-16 ***
PC3           1696.1       26.8   63.36  < 2e-16 ***
PC4          -1450.1       30.6  -47.39  < 2e-16 ***
PC5            881.4       32.2   27.40  < 2e-16 ***
PC6            956.0       34.1   28.07  < 2e-16 ***
PC7          -1528.2       38.9  -39.24  < 2e-16 ***
PC8           2544.8       40.3   63.16  < 2e-16 ***
PC9          -1230.3       43.5  -28.28  < 2e-16 ***
PC10          1155.2       47.0   24.56  < 2e-16 ***
PC11          2491.0      485.1    5.13  2.8e-07 ***
PC12         -1468.6    10240.7   -0.14     0.89
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8630 on 72080 degrees of freedom
Multiple R-squared:  0.666,     Adjusted R-squared:  0.665
F-statistic: 1.2e+04 on 12 and 72080 DF,  p-value: <2e-16
```

```
Call:
lm(formula = IncomePerCap ~ PC1, data = pcadata_pcr_rot)

Residuals:
   Min     1Q Median     3Q    Max
-65624  -5762  -1382   3924 199542

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28688.21      36.87   778.0   <2e-16 ***
PC1         -6049.05      19.99  -302.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9900 on 72091 degrees of freedom
Multiple R-squared:  0.5596,    Adjusted R-squared:  0.5596
F-statistic: 9.161e+04 on 1 and 72091 DF,  p-value: < 2.2e-16
```

K-means

K-means

Definition:

- Unsupervised clustering model for quantitative data

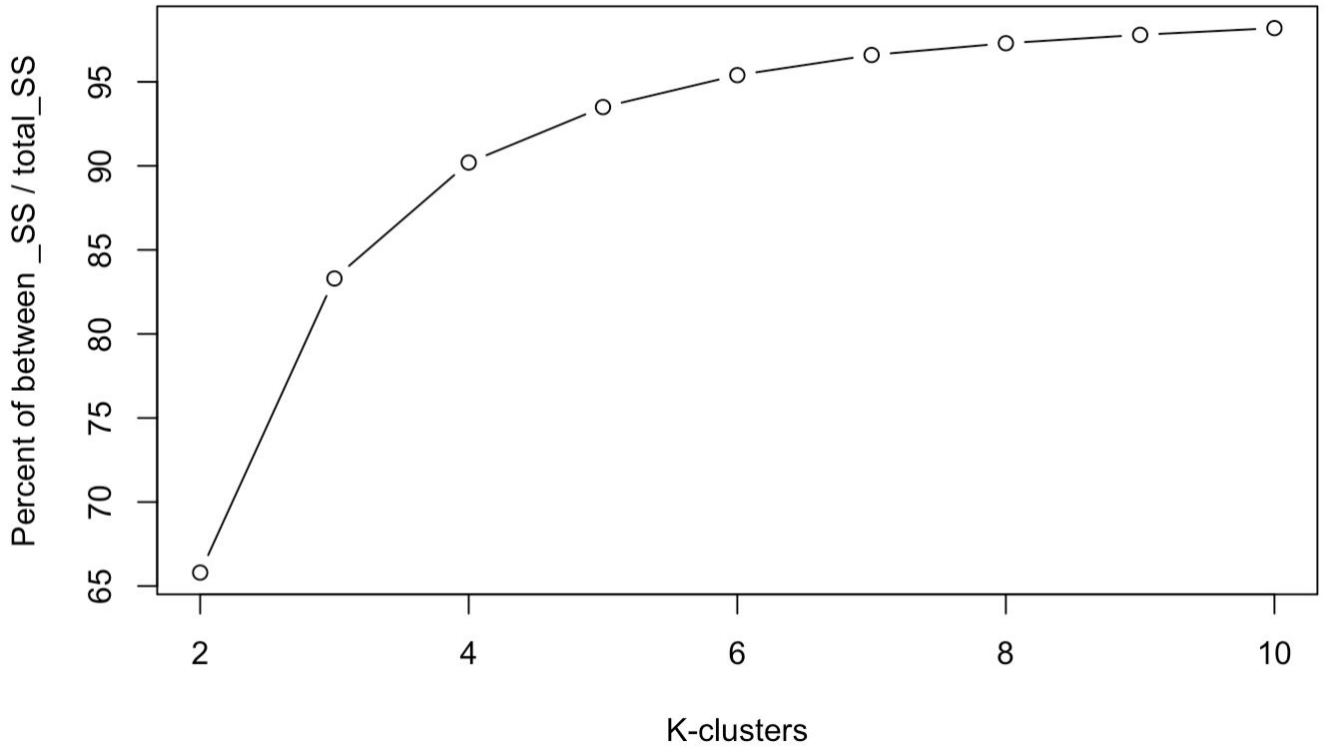
Pros:

- Very helpful in segmenting large datasets

Cons:

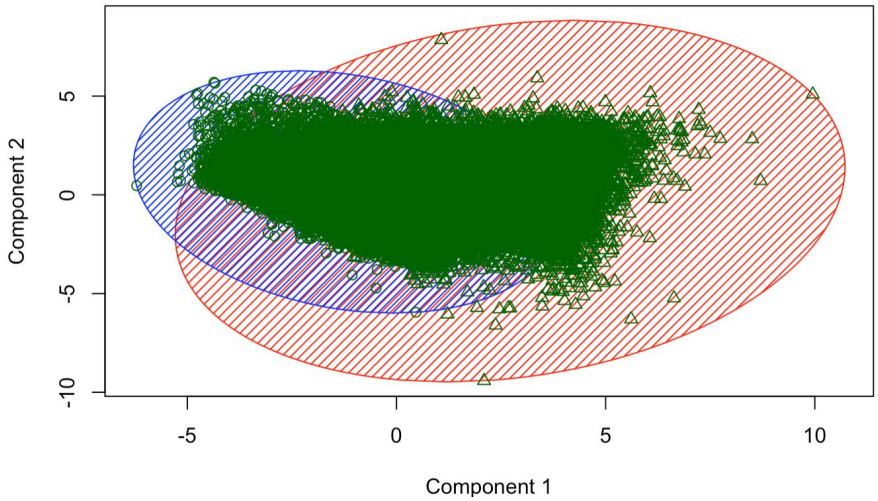
- Too many clusters will lead to smaller samples and may yield less useful information

K-means


Percent of between _SS / total_SS
65
70
75
80
85
90
95
2
4
6
8
10
K-clusters

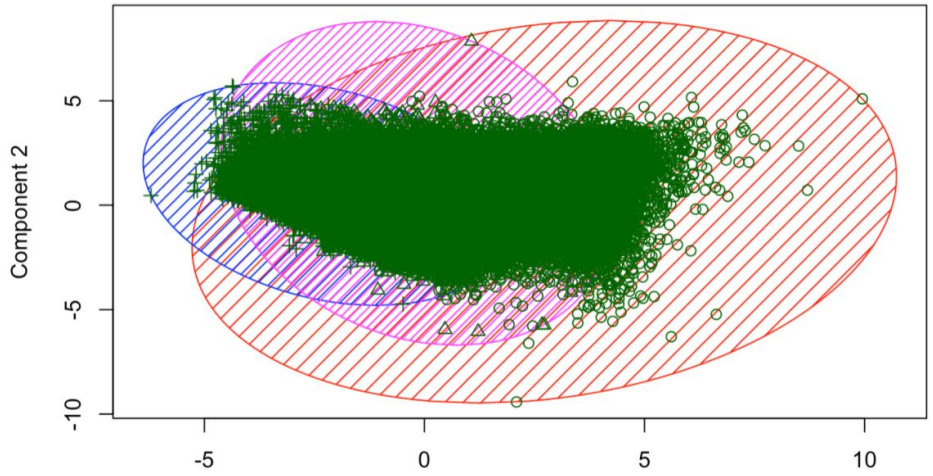
K-means (2)

```
Cluster means:
  Hispanic  White  Black  Asian Professional Service  Office Construction Production
1   -0.329  0.420 -0.320  0.237        0.928 -0.615 -0.0189       -0.430     -0.656
2    0.174 -0.222  0.169 -0.126       -0.492  0.326  0.0100        0.228      0.347
  Unemployment IncomePerCap
1       -0.527        37598
2        0.279        20114
```



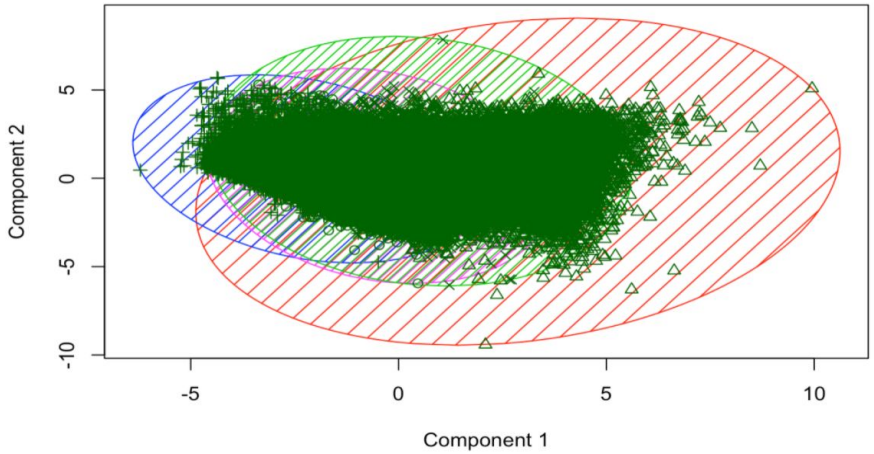
K-means (3)

```
Cluster means:
  Hispanic  White  Black   Asian Professional Service  Office Construction Production
1    0.432 -0.575  0.395 -0.1538       -0.769   0.613 -0.0261       0.2974      0.5120
2   -0.240  0.340 -0.208 -0.0241        0.133  -0.211  0.0733       0.0108     -0.0789
3   -0.363  0.435 -0.360  0.3837        1.330  -0.815 -0.1148      -0.6589     -0.9076
  Unemployment IncomePerCap
1        0.619        16577
2       -0.310        27822
3       -0.599        42760
```



K-means (4)

```
Cluster means:
  Hispanic  White   Black  Asian Professional Service  Office Construction Production
1   -0.302  0.407 -0.2817  0.107        0.571 -0.4351  0.0663       -0.229     -0.432
2    0.668 -0.881  0.5759 -0.165       -0.929  0.8325 -0.0548        0.314      0.575
3   -0.374  0.439 -0.3798  0.456        1.533 -0.9244 -0.1795       -0.769     -1.018
4   -0.132  0.185 -0.0808 -0.105       -0.241  0.0213  0.0495        0.186      0.224
  Unemployment IncomePerCap
1      -0.4610        32840
2       0.8988        14434
3      -0.6297        45781
4      -0.0999        23413
```



K—Nearest Neighbor

K—Nearest Neighbor

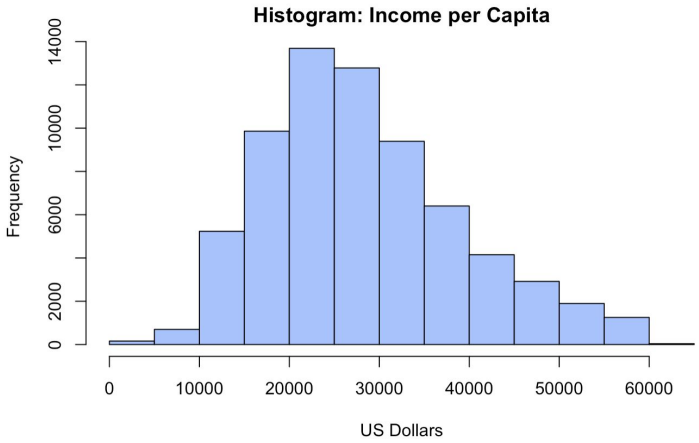
Advantages

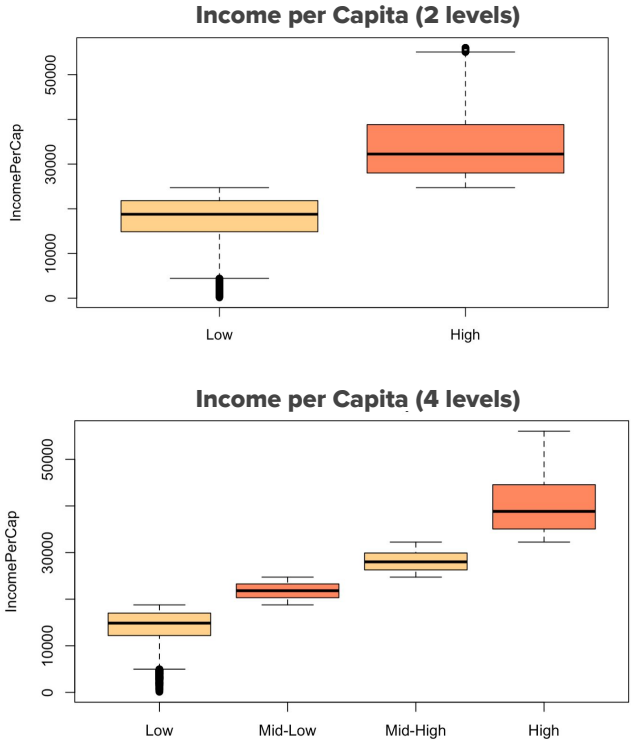
- Supervised learning model
- Classification (IncomePerCap)
- Several numeric features

Disadvantages

- IncomePerCap is numerical
- Inefficient since the entire training data is processed for every prediction

K—Nearest Neighbor





Low: 0 - 24,700

High: 24,700 - 56,000

Low: 0 - 18,000

Mid-Low: 18,800 - 24,700

Mid-High: 24,700 - 32,200

High: 32,000 - 56,000

K—Nearest Neighbor

Income per Capita (2 levels)

Predicted

True		Low	High
	Low	9480	1688
	High	1938	9730

Accuracy: 0.84
19,210 out of 22,836
K = 9

Income per Capita (4 levels)

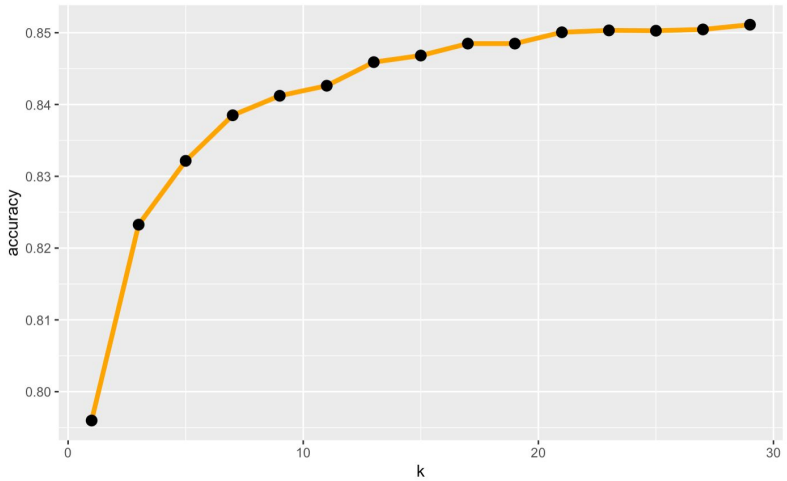
Predicted

True		Low	Mid-Low	Mid-High	High
	Low	4112	990	159	19
	Mid-Low	1287	3018	1436	159
	Mid-High	241	1487	2926	1089
	High	60	223	1220	4410

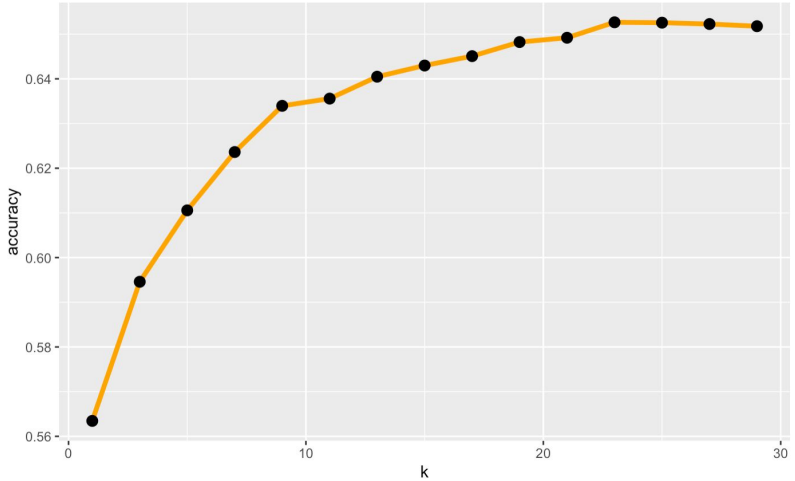
Accuracy: 0.63
14,466 out of 22,836
K = 9

K—Nearest Neighbor

Income per Capita (2 levels)

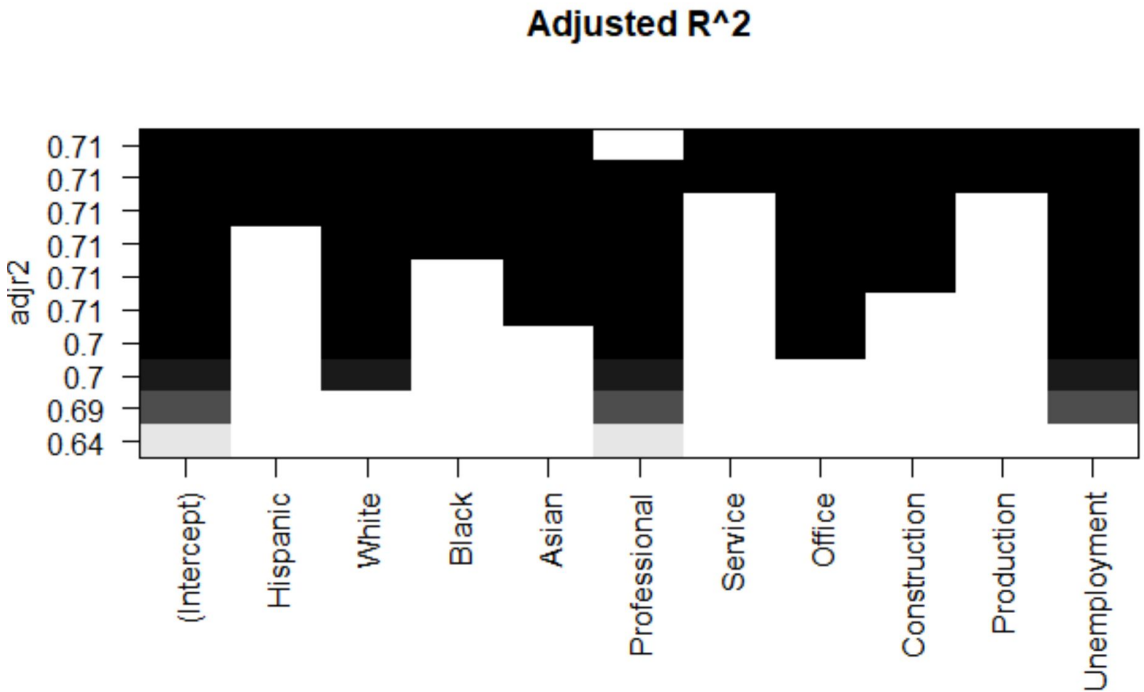


Income per Capita (4 levels)



Simple Regression

Exhaustive Search



Linear Regression

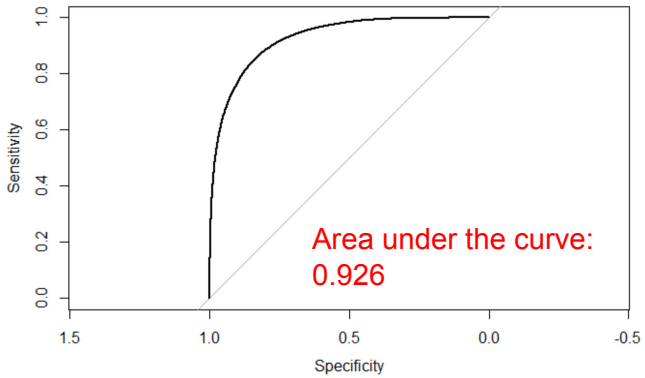
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  51988.355    396.421  131.14   <2e-16 ***
Hispanic        41.671      3.749   11.11   <2e-16 ***
White           96.026      3.777   25.42   <2e-16 ***
Black           53.109      3.803   13.97   <2e-16 ***
Asian          128.931      4.808   26.82   <2e-16 ***
Service       -536.921      3.174 -169.15   <2e-16 ***
Office        -383.300      3.848  -99.62   <2e-16 ***
Construction  -444.860      4.003 -111.12   <2e-16 ***
Production    -533.878      3.054 -174.82   <2e-16 ***
Unemployment  -270.607      4.495  -60.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5519 on 69557 degrees of freedom
  (105 observations deleted due to missingness)
Multiple R-squared:  0.7102,    Adjusted R-squared:  0.7101
F-statistic: 1.894e+04 on 9 and 69557 DF,  p-value: < 2.2e-16
```

- Positive effect from common ethnic groups
- Negative effects from non-professional employment.
- High significance
- High predictive power (R^2 = .7101)

Generalized Linear Model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   8.239739   0.274481  30.019  < 2e-16 ***
Hispanic      0.012491   0.002613   4.781 1.75e-06 ***
White         0.035533   0.002590  13.720  < 2e-16 ***
Black         0.016182   0.002654   6.096 1.09e-09 ***
Asian         0.050027   0.003193  15.668  < 2e-16 ***
Service      -0.178458   0.002239 -79.689  < 2e-16 ***
Office       -0.105114   0.002349 -44.744  < 2e-16 ***
Construction -0.134631   0.002510 -53.629  < 2e-16 ***
Production   -0.190259   0.002139 -88.936  < 2e-16 ***
Unemployment -0.140857   0.003127 -45.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 96440  on 69566  degrees of freedom
Residual deviance: 48210  on 69557  degrees of freedom
AIC: 48230

Number of Fisher Scoring iterations: 6
```

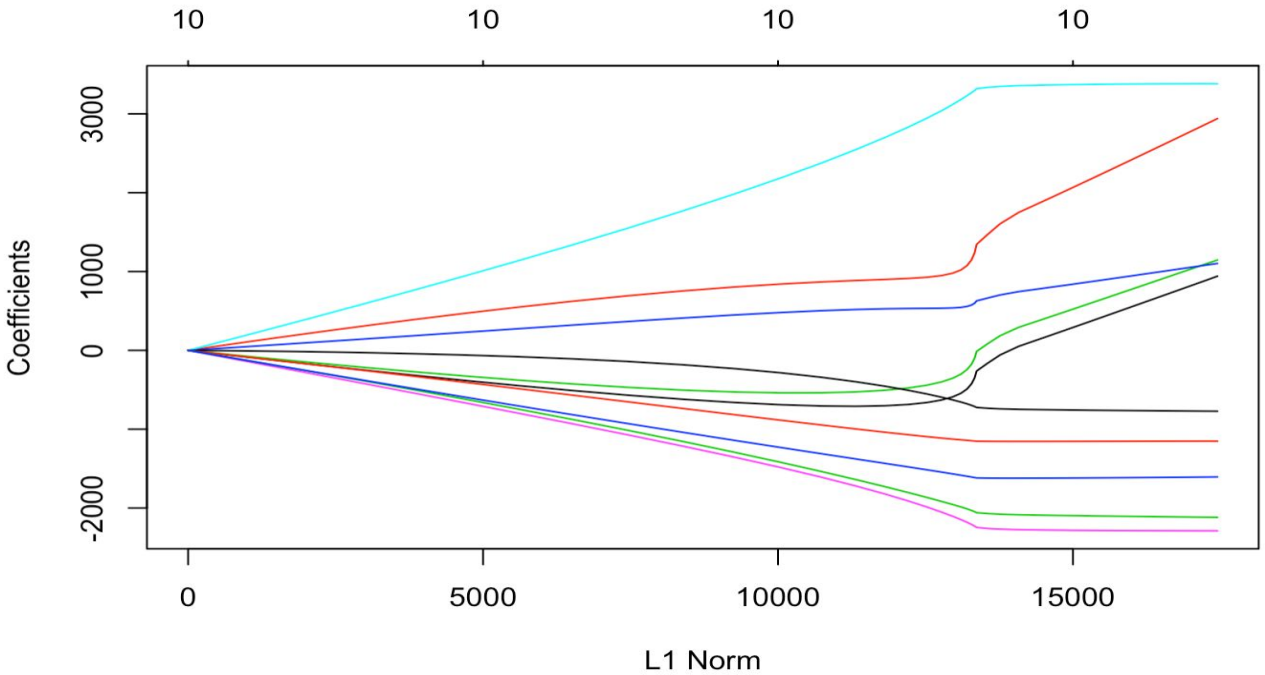
- Same effect directions and relative sizes
- More clearly see that higher white and asian populations increases the likelihood of falling into the wealthier half of the set of census tracts



Ridge Regression

Ridge Regression

```
All the coefficients :
 (Intercept)     Hispanic        White        Black        Asian Professional      Service
       26168         -463         1104         -215          564         3248        -2195
      Office Construction   Production Unemployment
        -688        -1144        -2021        -1602
```



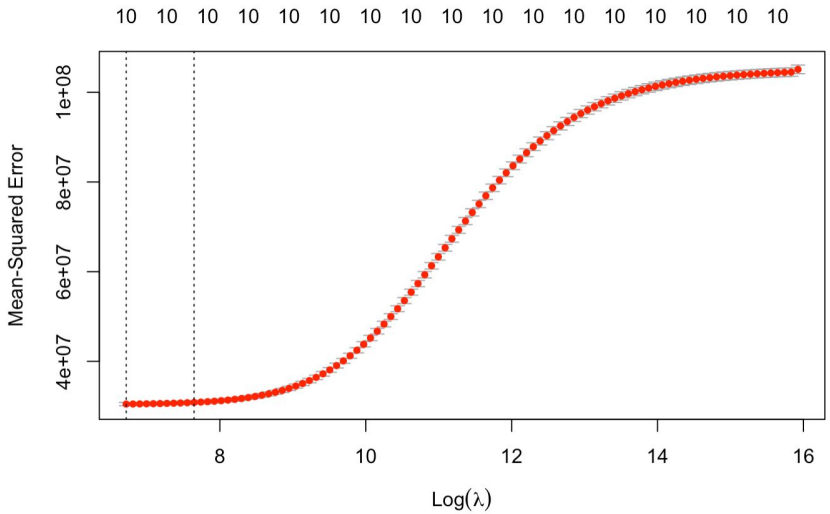
<Professional
<White

Ridge Regression

All the coefficients :

(Intercept)	Hispanic	White	Black	Asian	Professional	Service
26168	-463	1104	-215	564	3248	-2195

Office	Construction	Production	Unemployment
-688	-1144	-2021	-1602

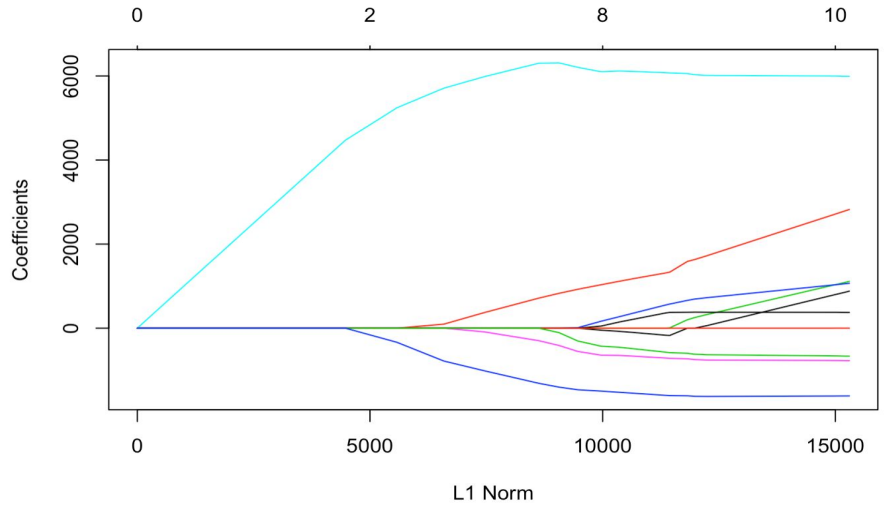


lowest lamda from CV: 825

MSE for best Ridge lamda: 30834392

Lasso Regression

```
The non-zero coefficients :
 (Intercept)     Hispanic        White        Black        Asian Professional      Service
     26167.8         13.6       1690.0        247.0        712.2       6030.2       -716.5
      Office   Production Unemployment
       367.5       -622.3      -1613.0
```

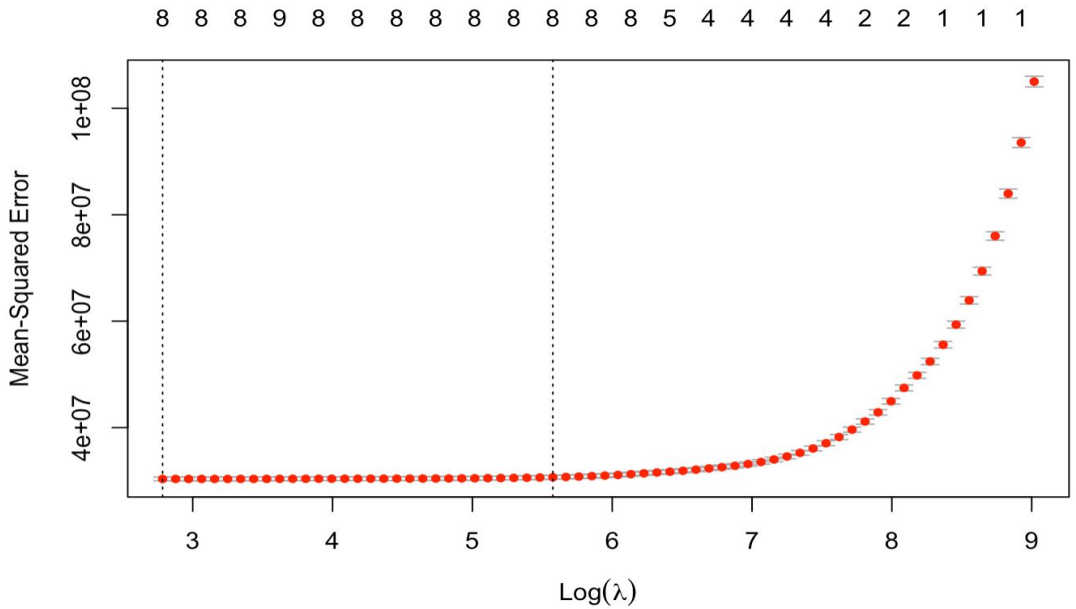


<Professional

<Unemployment

Lasso Regression continued

```
The non-zero coefficients :
 (Intercept)     Hispanic        White        Black        Asian Professional      Service
     26167.8         13.6       1690.0        247.0        712.2       6030.2       -716.5
      Office   Production Unemployment
       367.5       -622.3      -1613.0
```




```
lowest lamda from CV:  16.2

 MSE for best Lasso lamda:  30709528
```

OLS Full and Recommended Lasso

Full model

```
Call:
lm(formula = IncomePerCap ~ ., data = datJLClean)

Residuals:
   Min     1Q Median     3Q    Max
-57889  -3154   -136   3093  39355

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   26167.8       20.9 1250.46   <2e-16 ***
Hispanic        973.2       87.6   11.11   <2e-16 ***
White          2983.3      117.4   25.42   <2e-16 ***
Black          1175.9       84.2   13.97   <2e-16 ***
Asian          1115.2       41.6   26.82   <2e-16 ***
Professional    921.4     4378.8    0.21     0.83
Service       -3752.4     2603.4   -1.44     0.15
Office        -1839.0     1898.4   -0.97     0.33
Construction  -2236.0     1930.9   -1.16     0.25
Production    -3487.9     2435.8   -1.43     0.15
Unemployment  -1604.3       26.7  -60.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5520 on 69556 degrees of freedom
Multiple R-squared:  0.71,      Adjusted R-squared:  0.71
F-statistic: 1.7e+04 on 10 and 69556 DF,  p-value: <2e-16
```

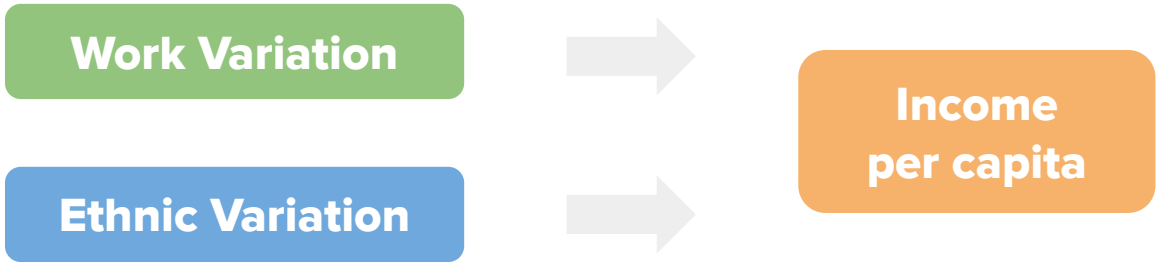
Without Construction

```
Call:
lm(formula = IncomePerCap ~ Hispanic + White + Black + Asian +
    Professional + Service + Office + Production + Unemployment,
    data = datJLClean)

Residuals:
   Min     1Q Median     3Q    Max
-57889  -3155   -139   3092  39315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   26167.8       20.9  1250.5   <2e-16 ***
Hispanic        973.0       87.6    11.1   <2e-16 ***
White          2983.2      117.4    25.4   <2e-16 ***
Black          1175.9       84.2    14.0   <2e-16 ***
Asian          1115.2       41.6    26.8   <2e-16 ***
Professional   5991.9       53.9   111.1   <2e-16 ***
Service        -737.9       39.8   -18.6   <2e-16 ***
Office          359.1       28.6    12.6   <2e-16 ***
Production     -667.6       40.6   -16.4   <2e-16 ***
Unemployment  -1604.1       26.7   -60.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5520 on 69557 degrees of freedom
Multiple R-squared:  0.71,      Adjusted R-squared:  0.71
F-statistic: 1.89e+04 on 9 and 69557 DF,  p-value: <2e-16
```

Overview of Lasso, Ridge, and Selected OLS

Model	MSE	R^2
Ridge	30834392	.707
Lasso	30709528	.708
Full model	30459848	.71
Full model w/o construction	30460435	.71

Conclusion



Citations

MuonNeutrino. (2015). US Census Demographic Data: Demographic and Economic Data for Tracts and Counties. UpToDate. Retrieved March 23, 2020, from https://www.kaggle.com/muonneutrino/us-census-demographic-dataD

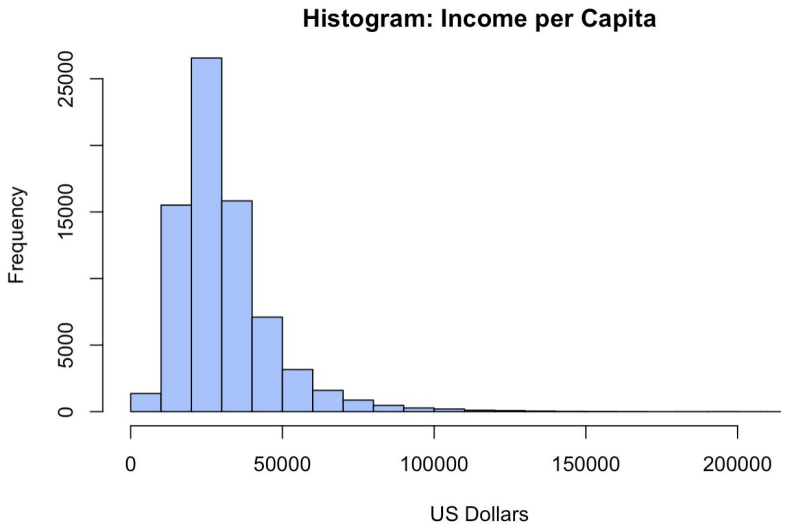
U.S. Census Bureau (2019). "Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019". 2010-2019 Population Estimates. United States Census Bureau, Population Division. December 30, 2019. Retrieved January 27, 2020.

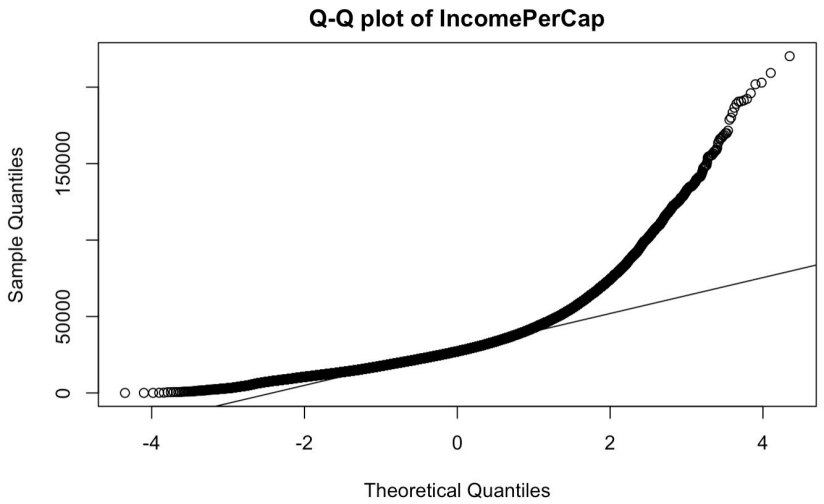
U.S. Census Bureau (2017). "American FactFinder - Results". U.S. Census Bureau. Retrieved 2017-12-13.

U.S. Census Bureau (2013). "2010 Census Summary File 1: GEOGRAPHIC IDENTIFIERS". American Factfinder. US Census. Retrieved 18 October 2013.

Appendix

Income per Capita

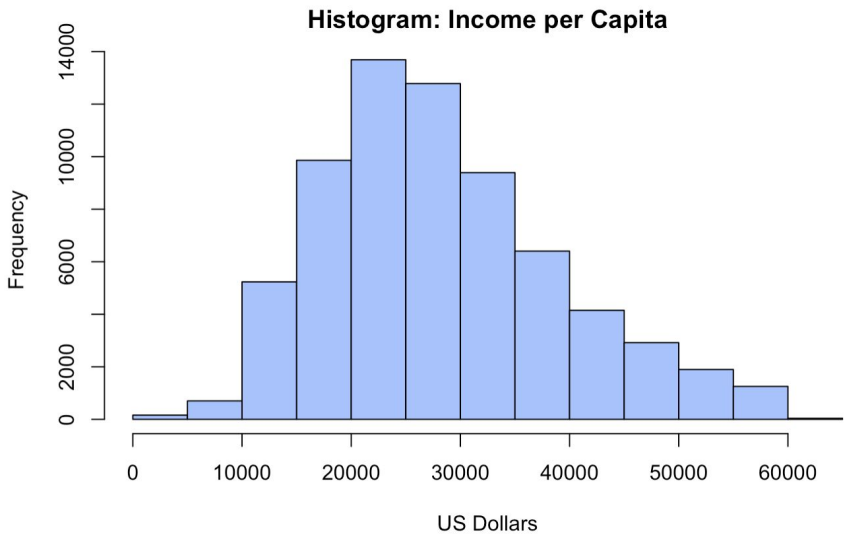


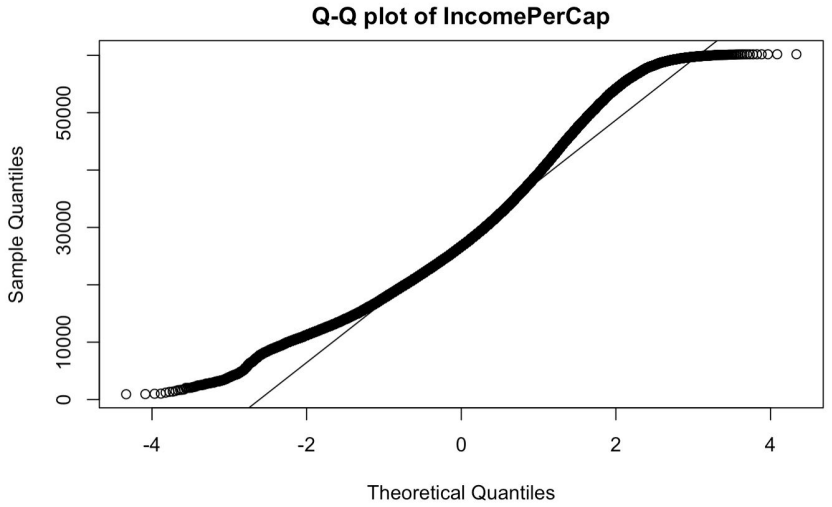


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
32	20557	27216	30652	36408	220253	745

Income per Capita

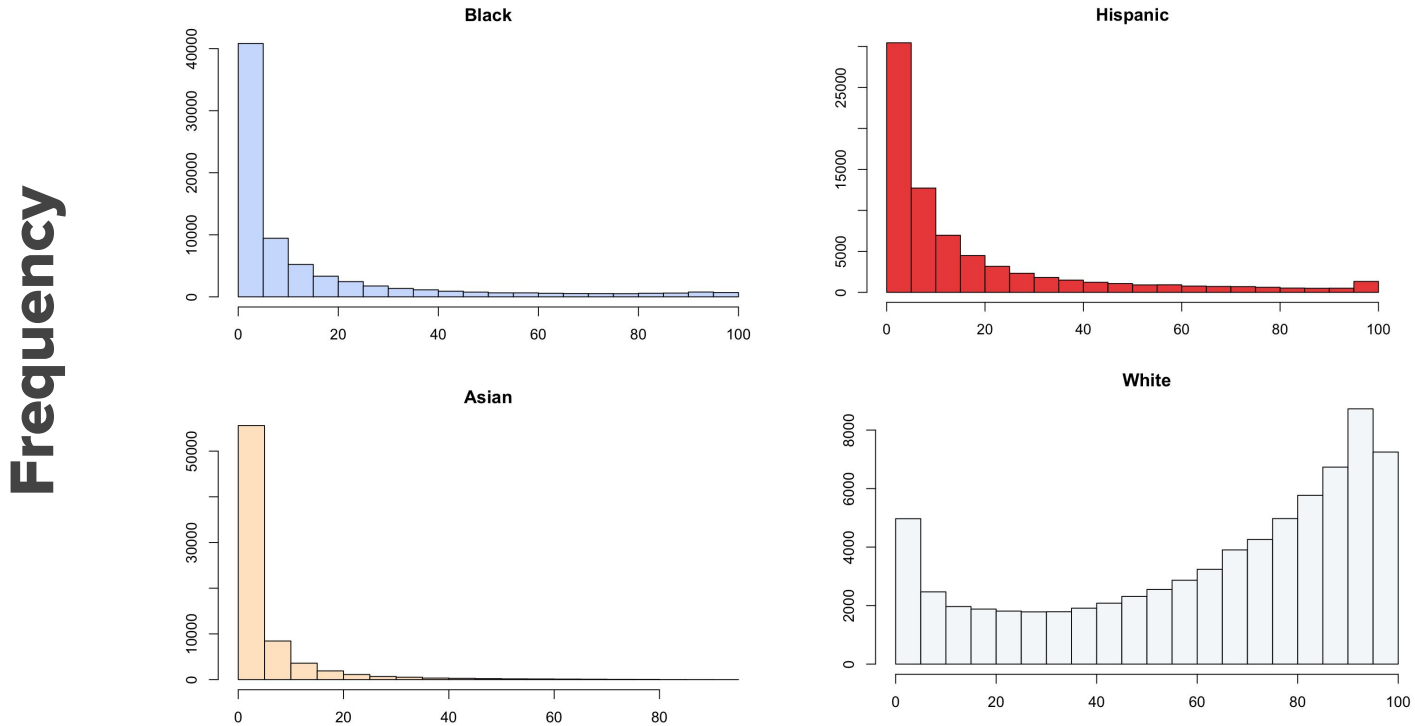
After removing Outliers and NA's



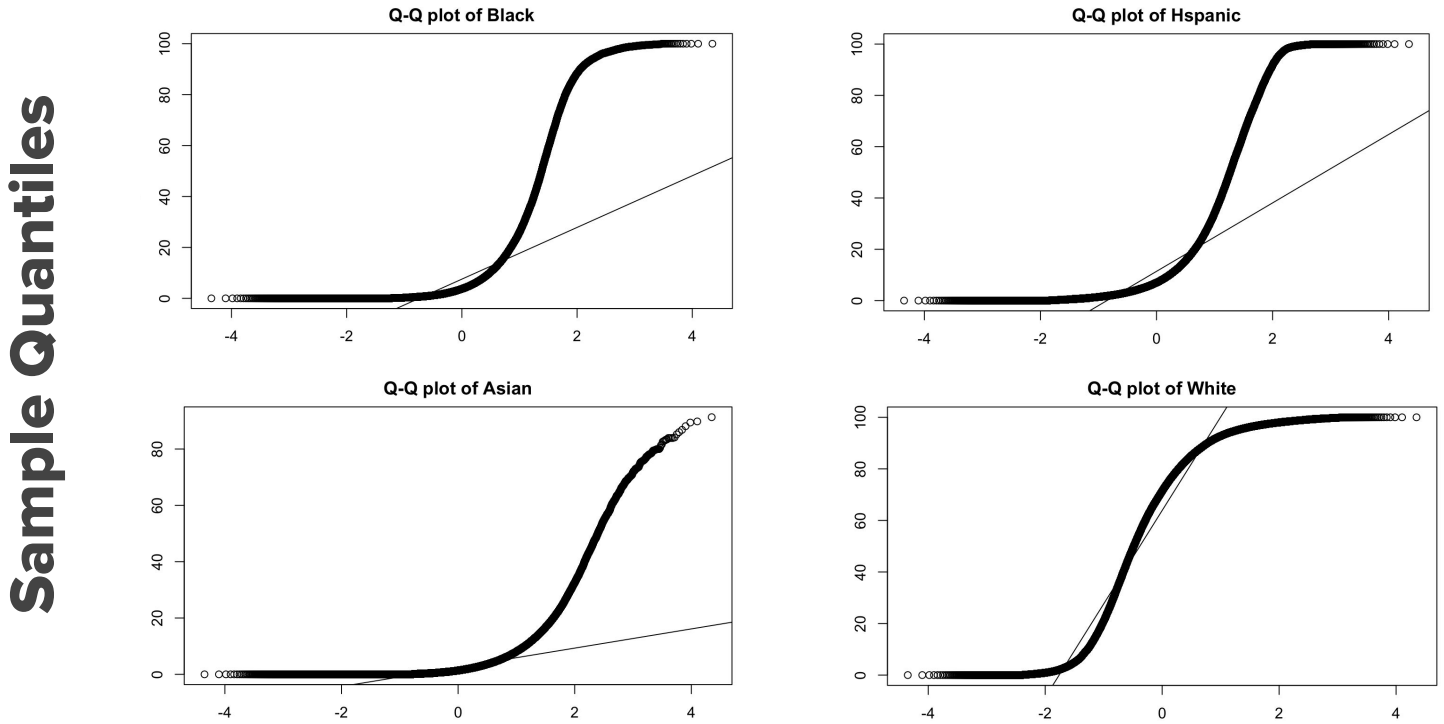



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 923   20453   26698   28309   34708   60185
```

Histogram Ethnic Variation

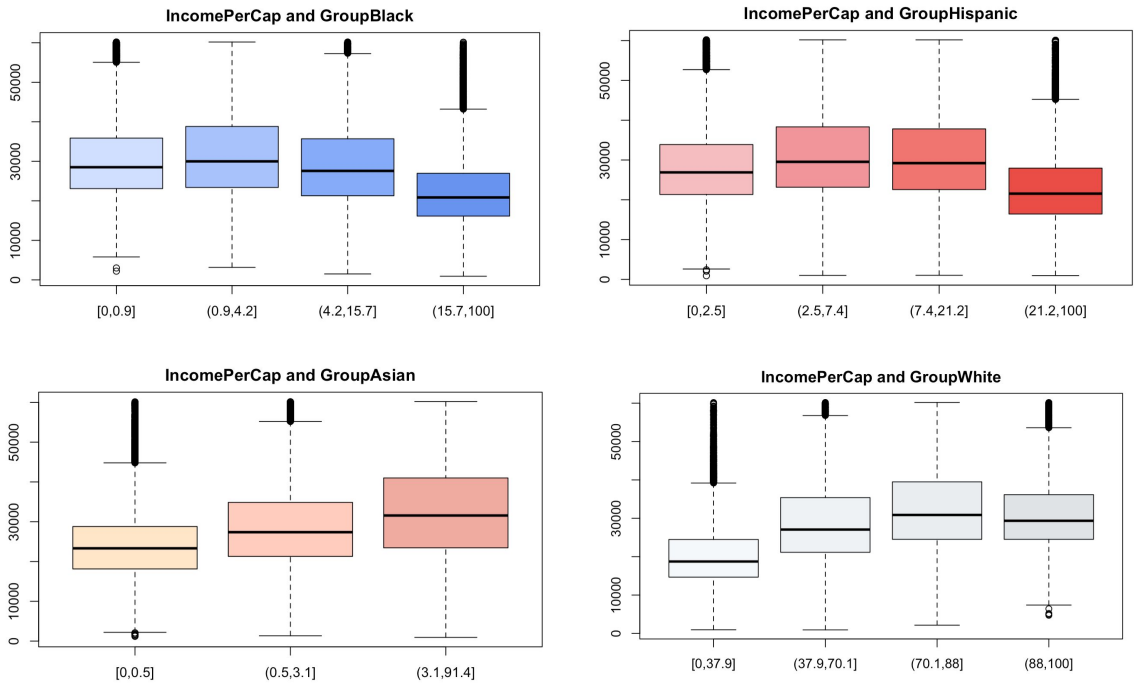


Q-Q plot Ethnic Variation



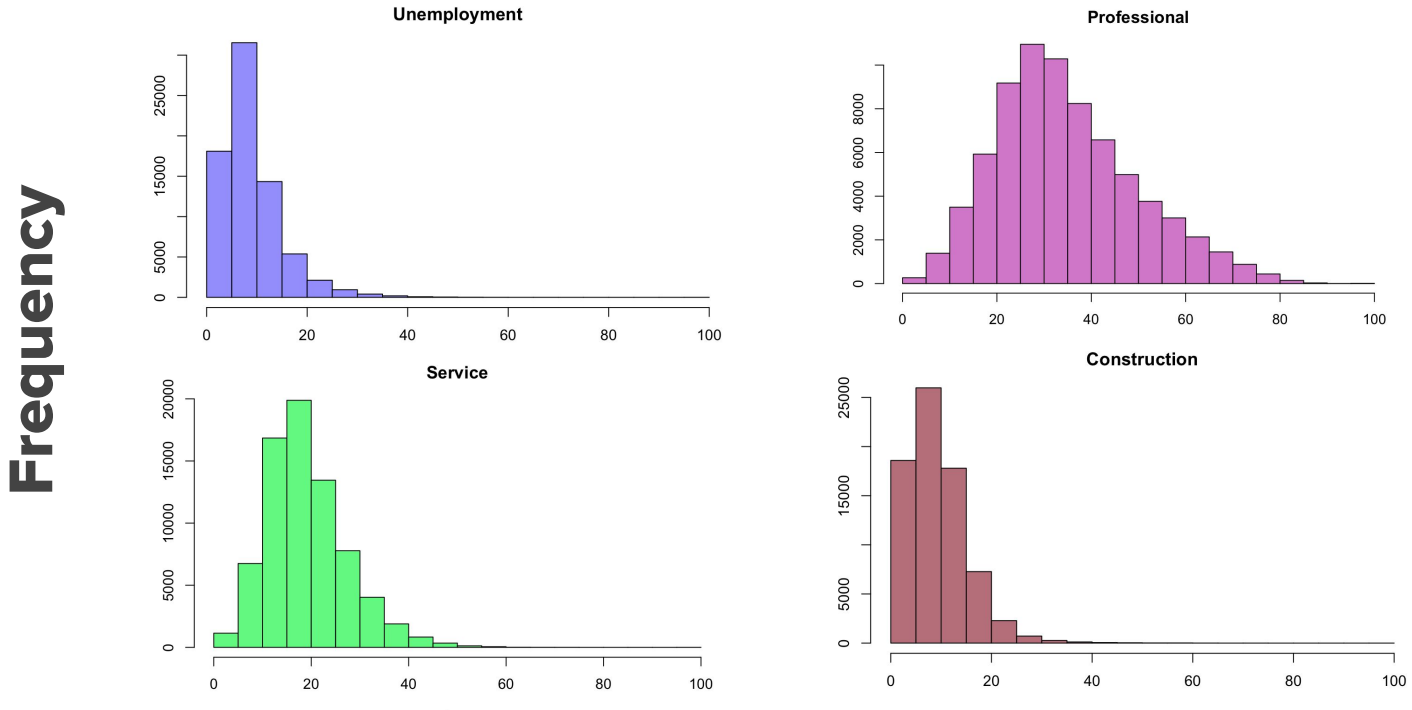
Boxplots Ethnic Variation

$US Dollars

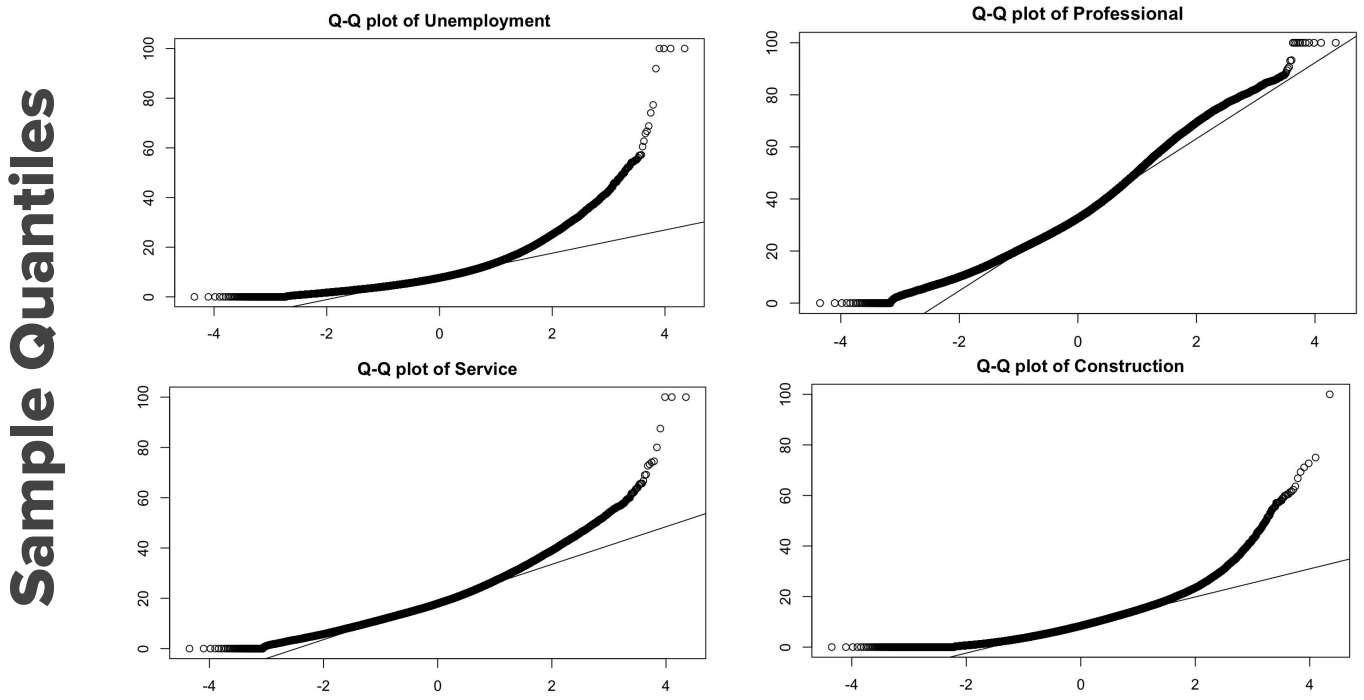


% in Census Tract

Histogram Work Variation

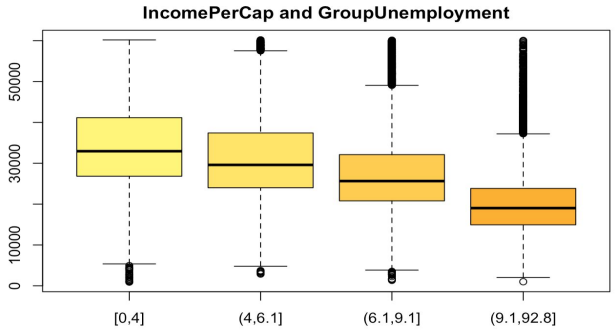


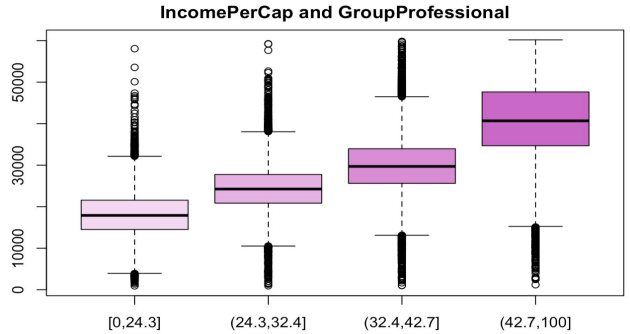
Q-Q plot Work Variation

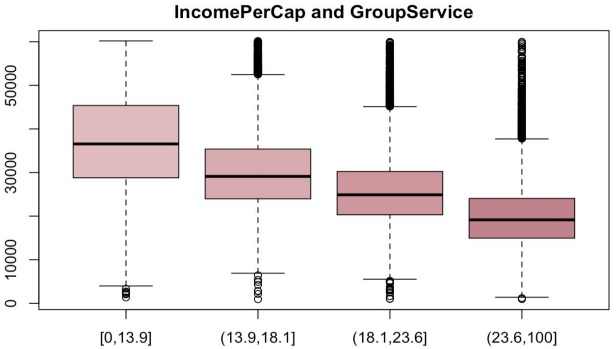


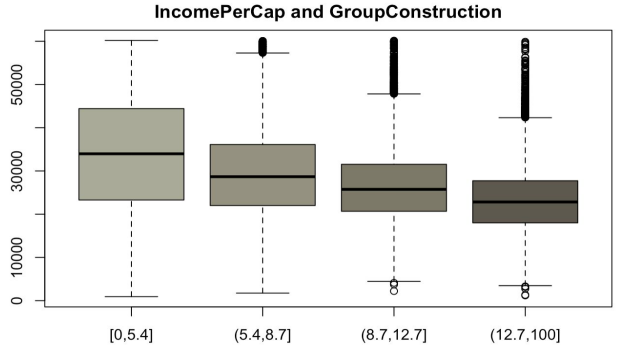
Boxplots Work Variation

$US Dollars









% in Census Tract