Creating R programs using Jupyter notebook

Megha Patel

Economist

Office of Price and Living Conditions

Acceptance Testing Team

Background

- Master’s Graduate in Economic Development - Vanderbilt University
  - Specialization: Economics of Poverty in Developed and Developing Countries, Microeconomics
- Undergraduate Degree: Economics – Rutgers University
- Skillset: R/RStudio, Python, STATA
  - Mapping, Data Visualizations, and Statistics

Presentation Flow

- What is Jupyter Notebook?
- Advantages for R Users
- Building A Data Science Project

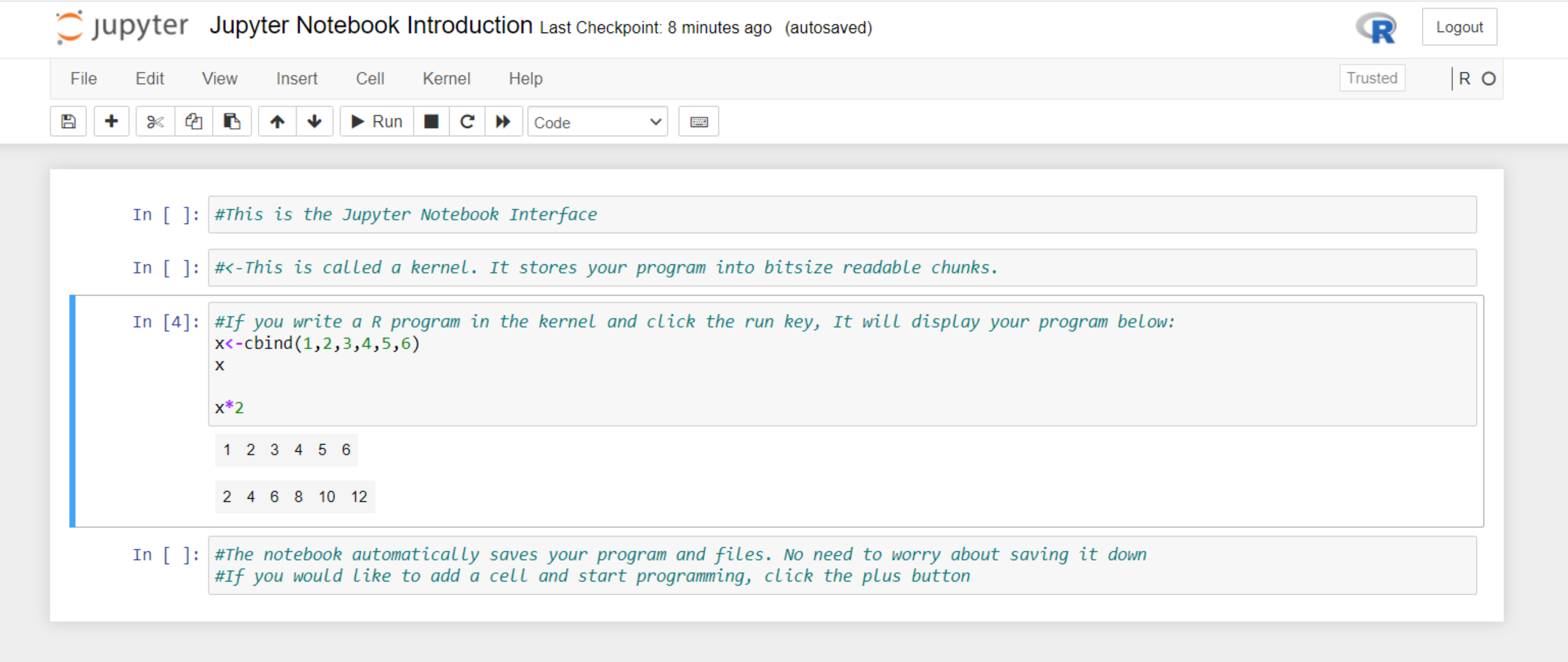
Jupyter Notebook

- "The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more."
  - https://jupyter.org/
  - Supports over 40 Languages

What's the Advantage for R Users?

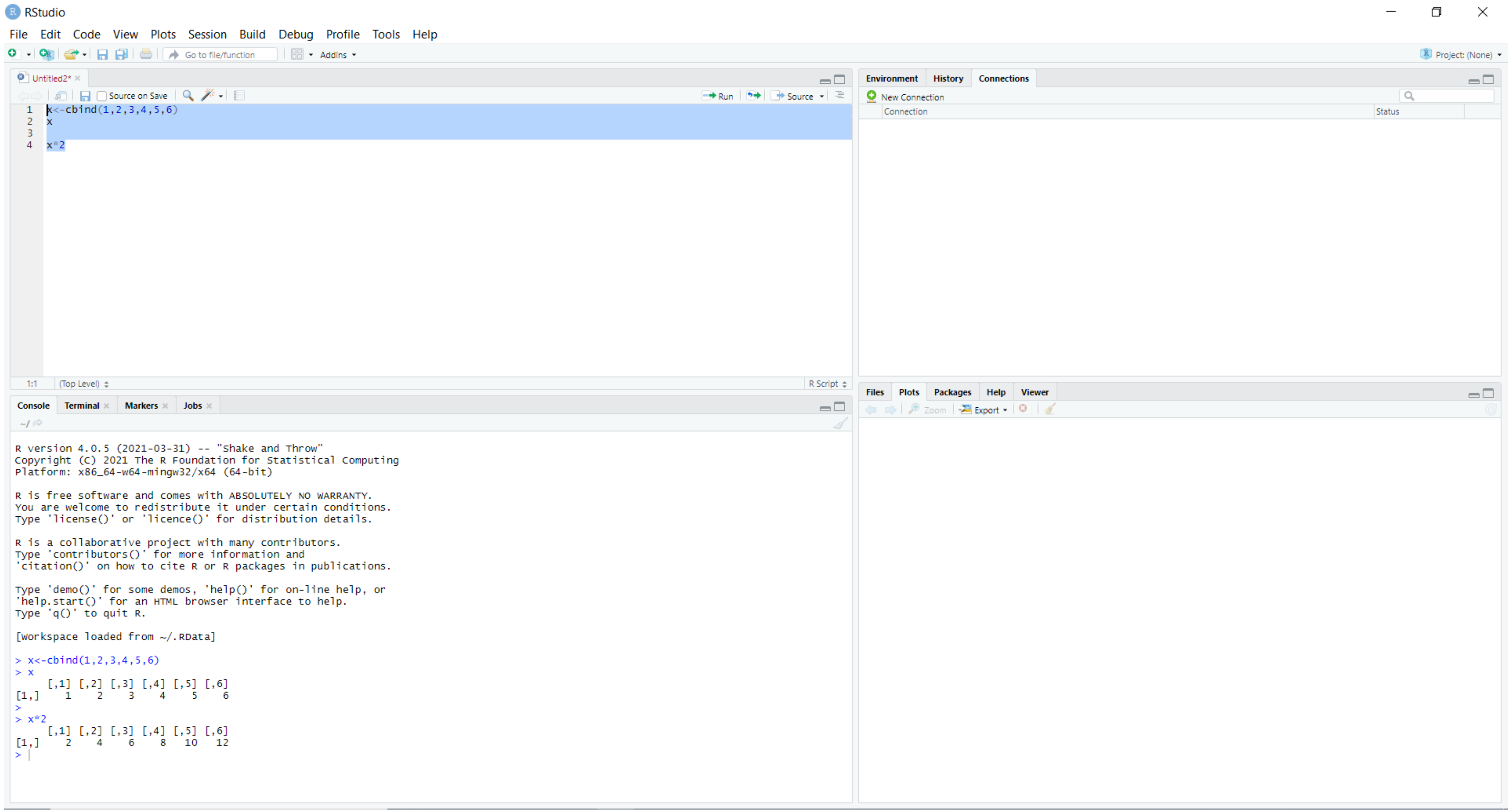
- Easy to Build Projects for Users
  - Interface makes it easier to connect different outputs with coding lines
- Organizing Complex Data Science Projects
  - R/SAS, R/SQL, R/Python, etc.
- Readability and Accessibility
  - File outputs are readable and interactive for audiences

Jupyter Notebook Interface


jupyter
Jupyter Notebook Introduction Last Checkpoint: 8 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Help
Trusted
R
Run
Code

```
In [ ]: #This is the Jupyter Notebook Interface

In [ ]: #<-This is called a kernel. It stores your program into bitsize readable chunks.

In [4]: #If you write a R program in the kernel and click the run key, It will display your program below:
        x<-cbind(1,2,3,4,5,6)
        x

        x*2

        1  2  3  4  5  6

        2  4  6  8  10  12

In [ ]: #The notebook automatically saves your program and files. No need to worry about saving it down
        #If you would like to add a cell and start programming, click the plus button
```

RStudio Interface



```r
x<-cbind(1,2,3,4,5,6)
x

x*2
```

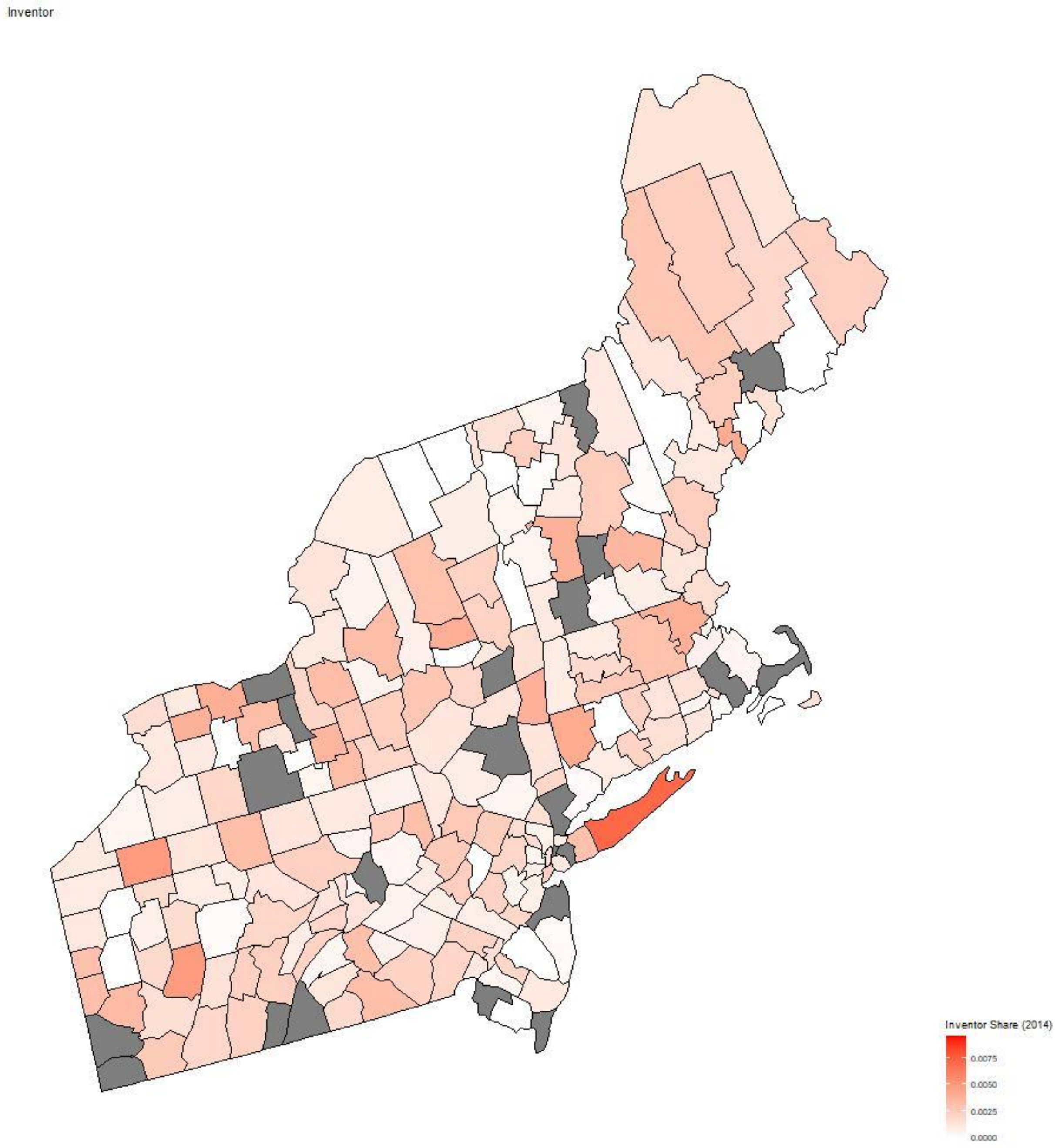
```
R version 4.0.5 (2021-03-31) -- "Shake and Throw"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> x<-cbind(1,2,3,4,5,6)
> x
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    2    3    4    5    6
>
> x*2
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    2    4    6    8   10   12
> 
```

Building A Data Science Project

- Social Origins of Inventors
  - Aim: Find birthplaces of engineers in the United States
  - Results: Low Fit of model suggests that neighborhood level characteristics are not as predictive for female engineers as they are for male engineers
  - Tools: R/RStudio and Jupyter Notebook
  - Data Source: Opportunity Insights
    - Raj Chetty, Harvard University

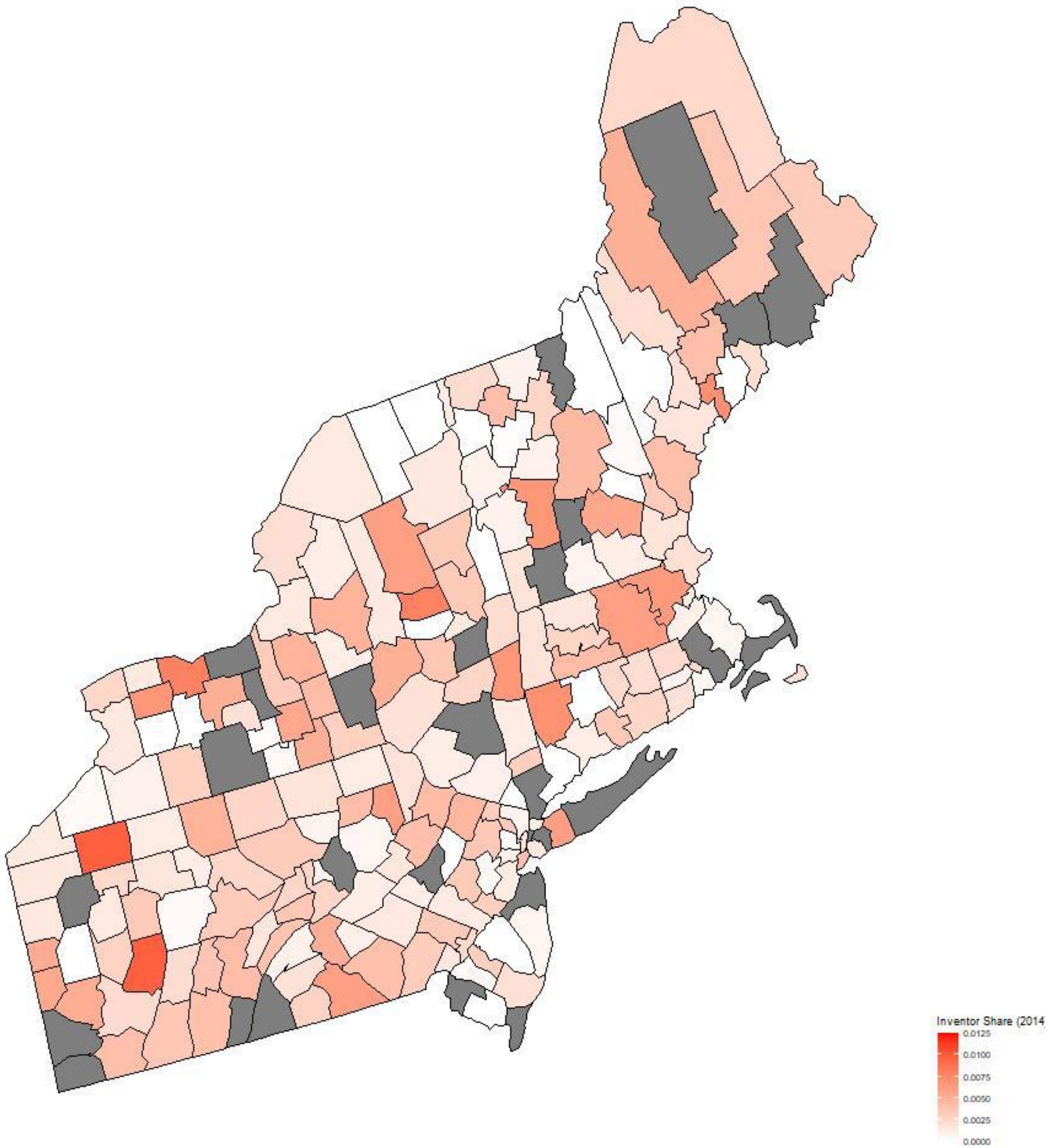
Demo

Results Using Jupyter Notebook

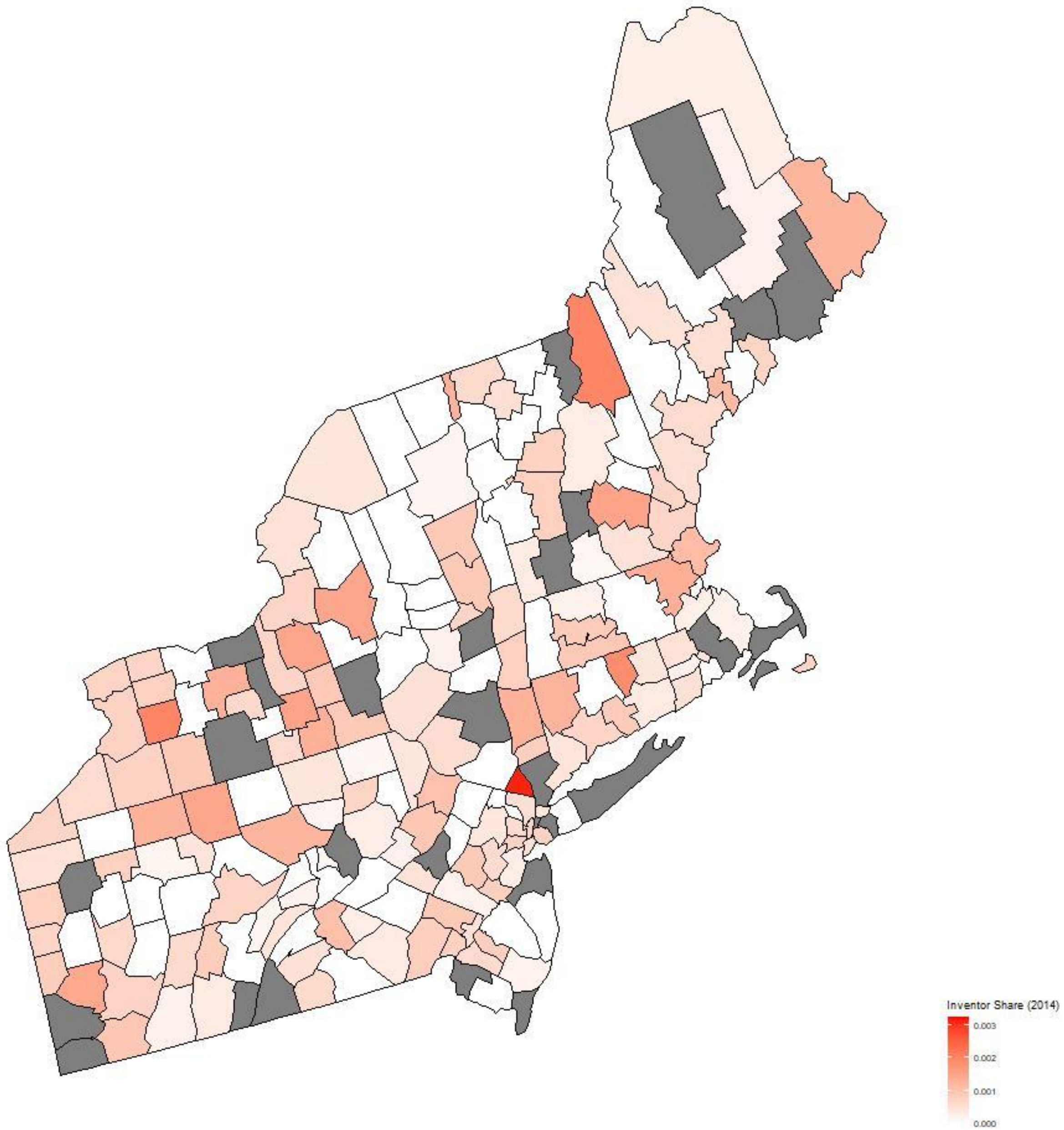


Plotting ggplot2 objects on the Jupyter Notebook Interface

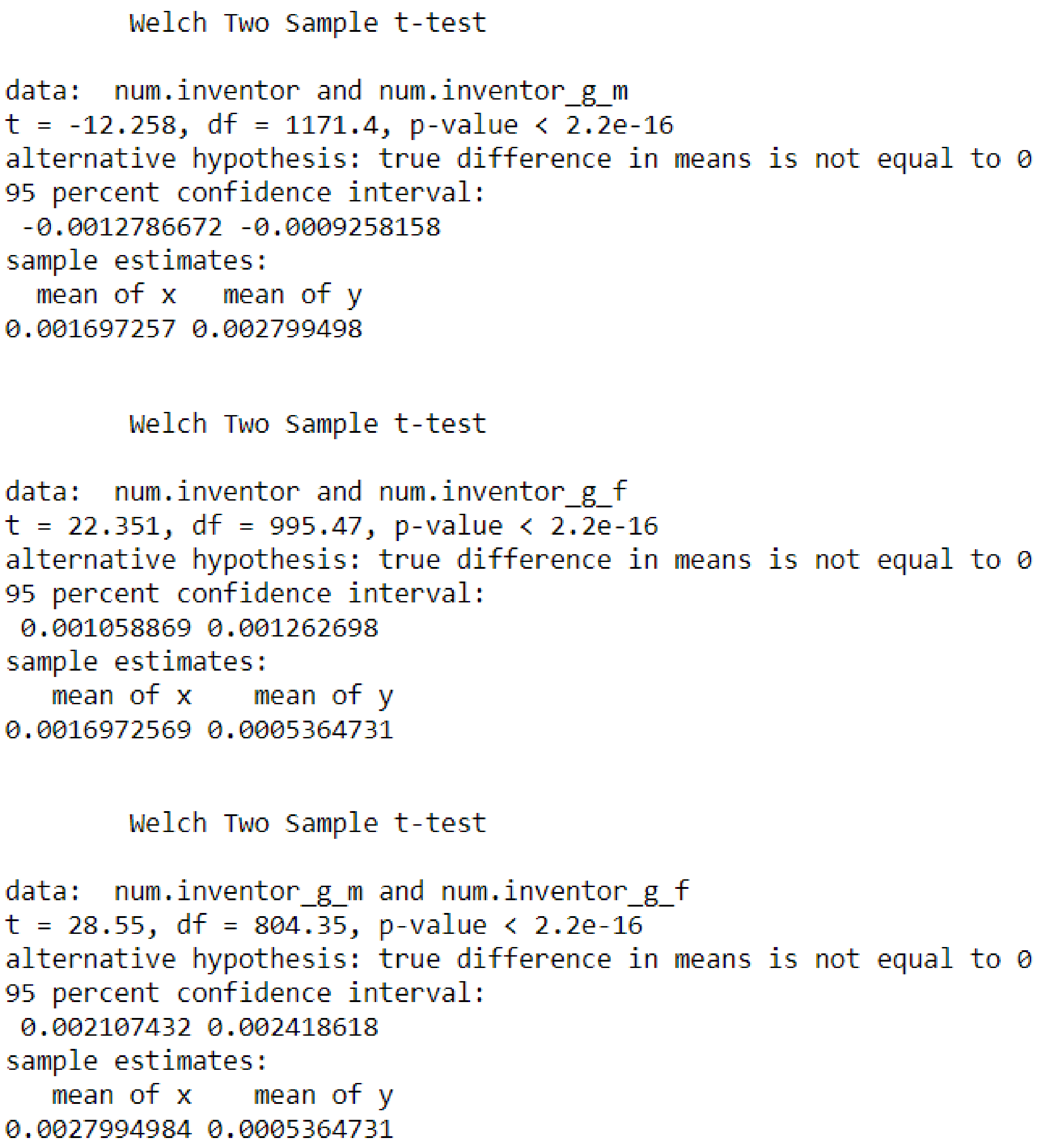
Male Inventor



Female Inventor

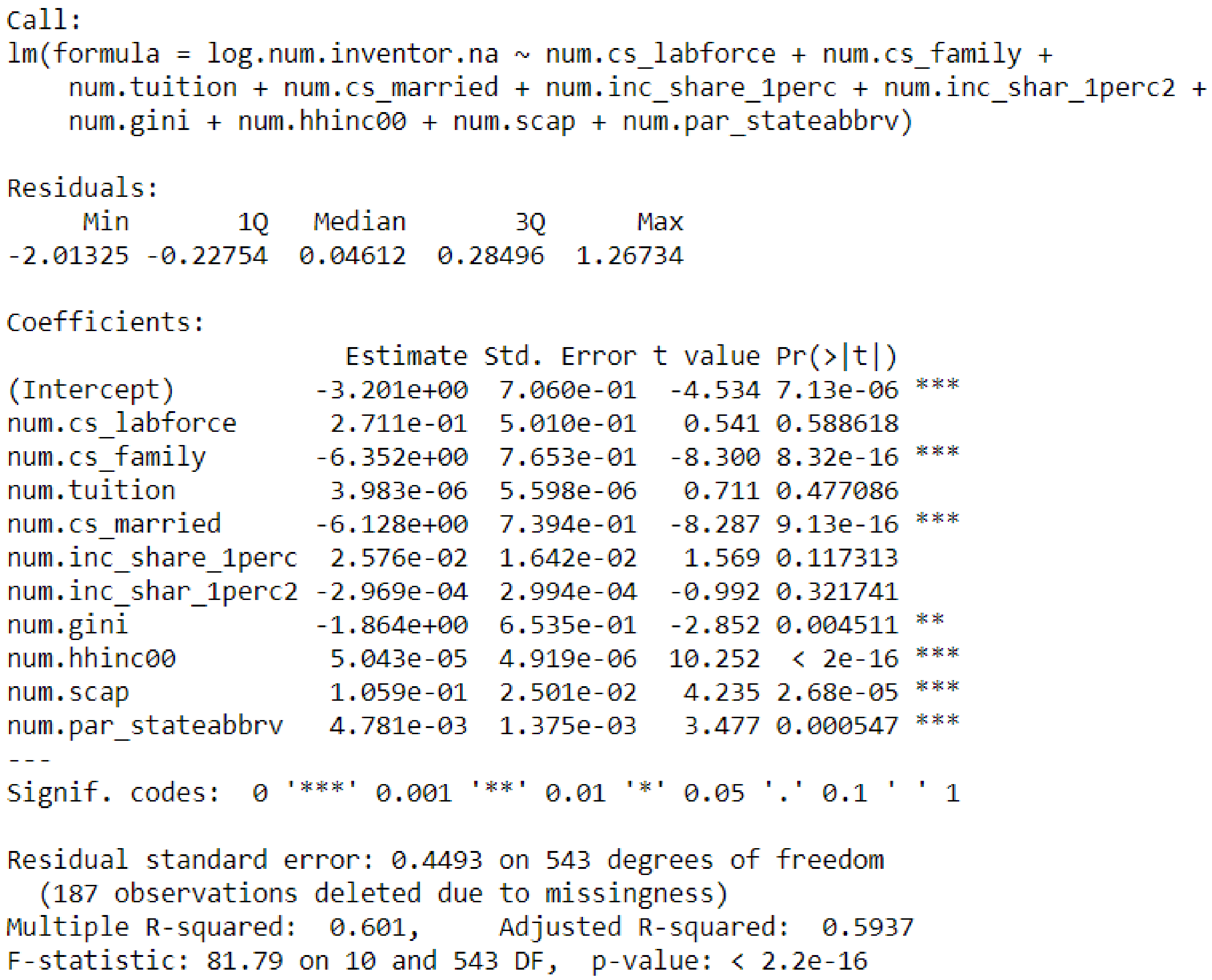


Outputs for Building T-Tests



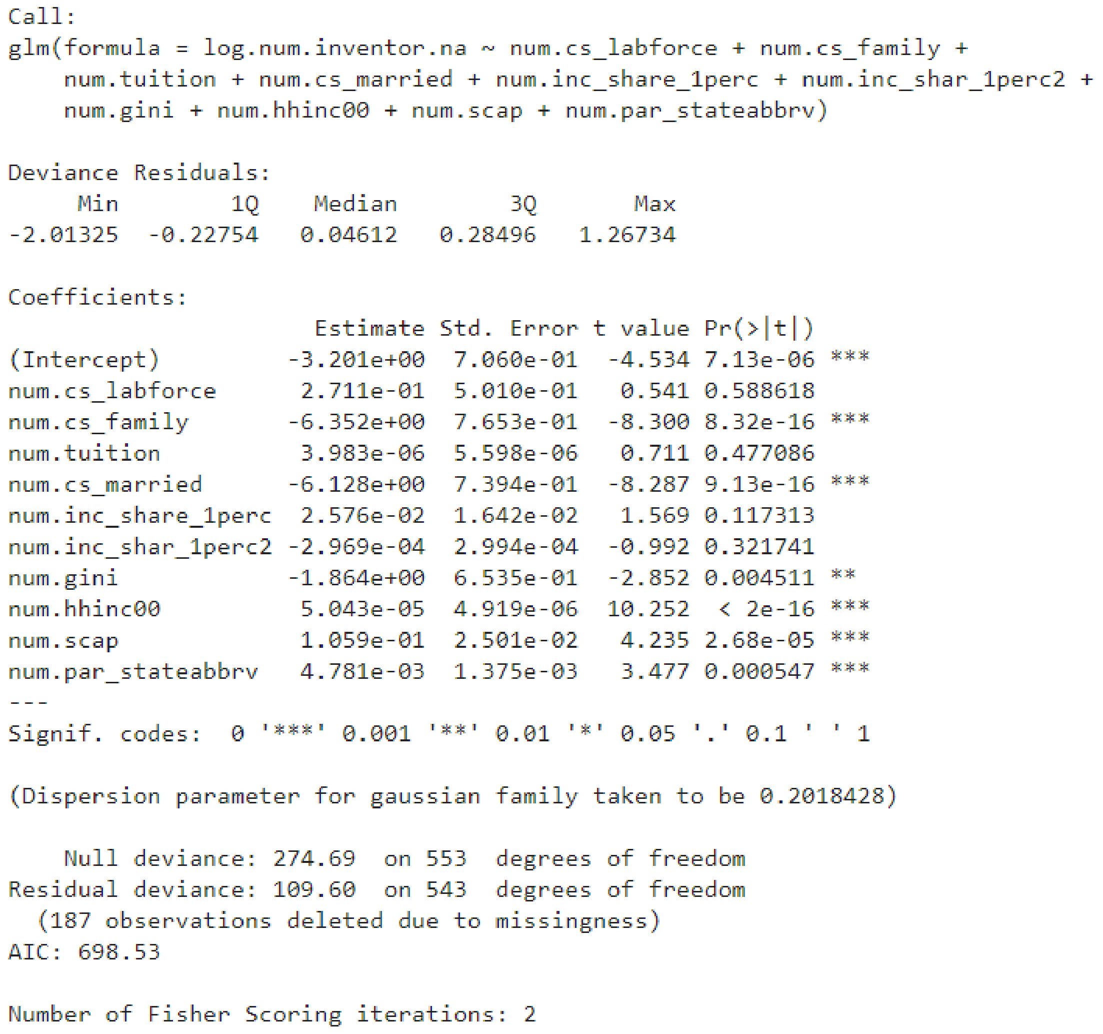
```
        Welch Two Sample t-test

data:  num.inventor and num.inventor_g_m
t = -12.258, df = 1171.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0012786672 -0.0009258158
sample estimates:
  mean of x   mean of y 
0.001697257 0.002799498 


        Welch Two Sample t-test

data:  num.inventor and num.inventor_g_f
t = 22.351, df = 995.47, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.001058869 0.001262698
sample estimates:
   mean of x    mean of y 
0.0016972569 0.0005364731 


        Welch Two Sample t-test

data:  num.inventor_g_m and num.inventor_g_f
t = 28.55, df = 804.35, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.002107432 0.002418618
sample estimates:
   mean of x    mean of y 
0.0027994984 0.0005364731 
```

lm(): Multiple Regression



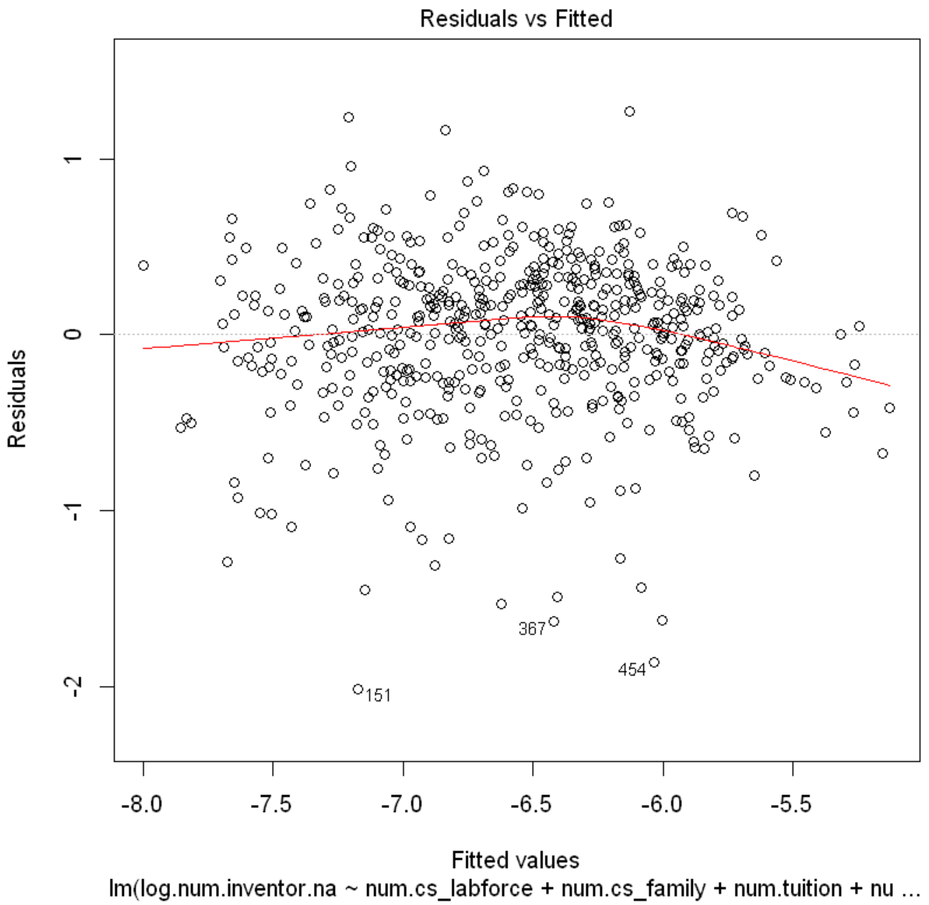
```
Call:
lm(formula = log.num.inventor.na ~ num.cs_labforce + num.cs_family + 
    num.tuition + num.cs_married + num.inc_share_1perc + num.inc_shar_1perc2 + 
    num.gini + num.hhinc00 + num.scap + num.par_stateabbrv)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.01325 -0.22754  0.04612  0.28496  1.26734 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -3.201e+00  7.060e-01  -4.534 7.13e-06 ***
num.cs_labforce      2.711e-01  5.010e-01   0.541 0.588618    
num.cs_family       -6.352e+00  7.653e-01  -8.300 8.32e-16 ***
num.tuition          3.983e-06  5.598e-06   0.711 0.477086    
num.cs_married      -6.128e+00  7.394e-01  -8.287 9.13e-16 ***
num.inc_share_1perc  2.576e-02  1.642e-02   1.569 0.117313    
num.inc_shar_1perc2 -2.969e-04  2.994e-04  -0.992 0.321741    
num.gini            -1.864e+00  6.535e-01  -2.852 0.004511 ** 
num.hhinc00          5.043e-05  4.919e-06  10.252  < 2e-16 ***
num.scap             1.059e-01  2.501e-02   4.235 2.68e-05 ***
num.par_stateabbrv   4.781e-03  1.375e-03   3.477 0.000547 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4493 on 543 degrees of freedom
  (187 observations deleted due to missingness)
Multiple R-squared:  0.601,	Adjusted R-squared:  0.5937 
F-statistic: 81.79 on 10 and 543 DF,  p-value: < 2.2e-16
```

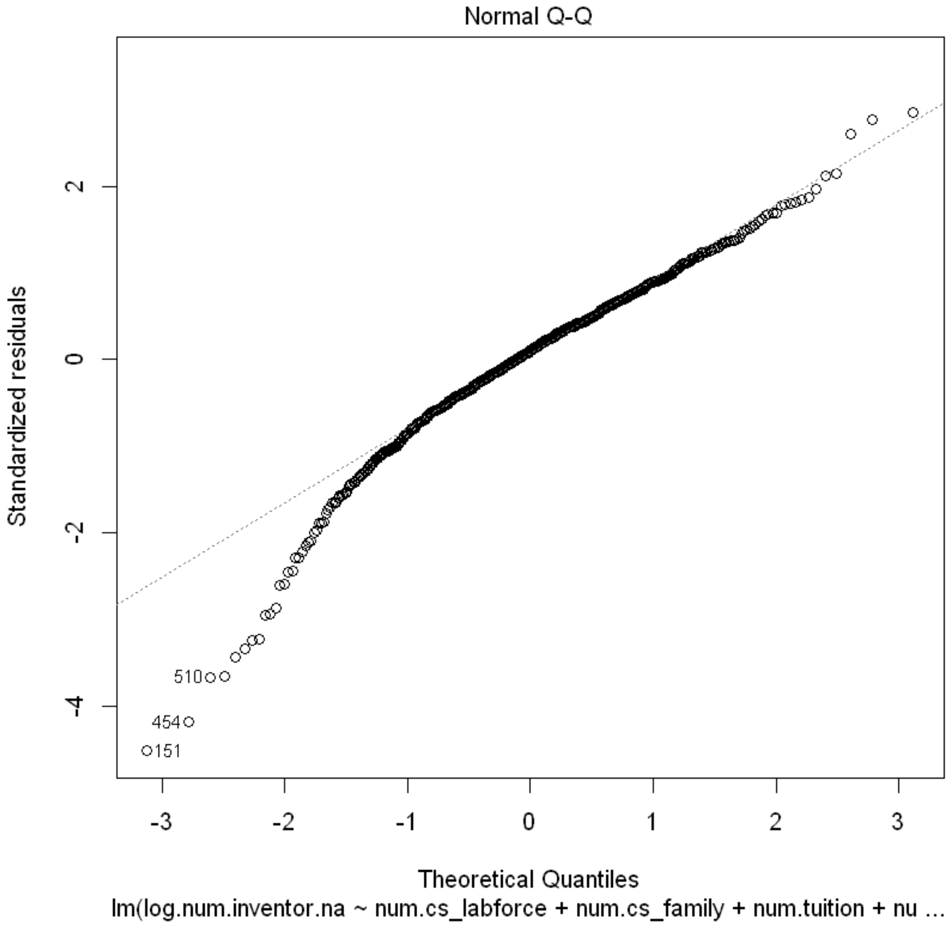
glm(): Gaussian Family Regression Models

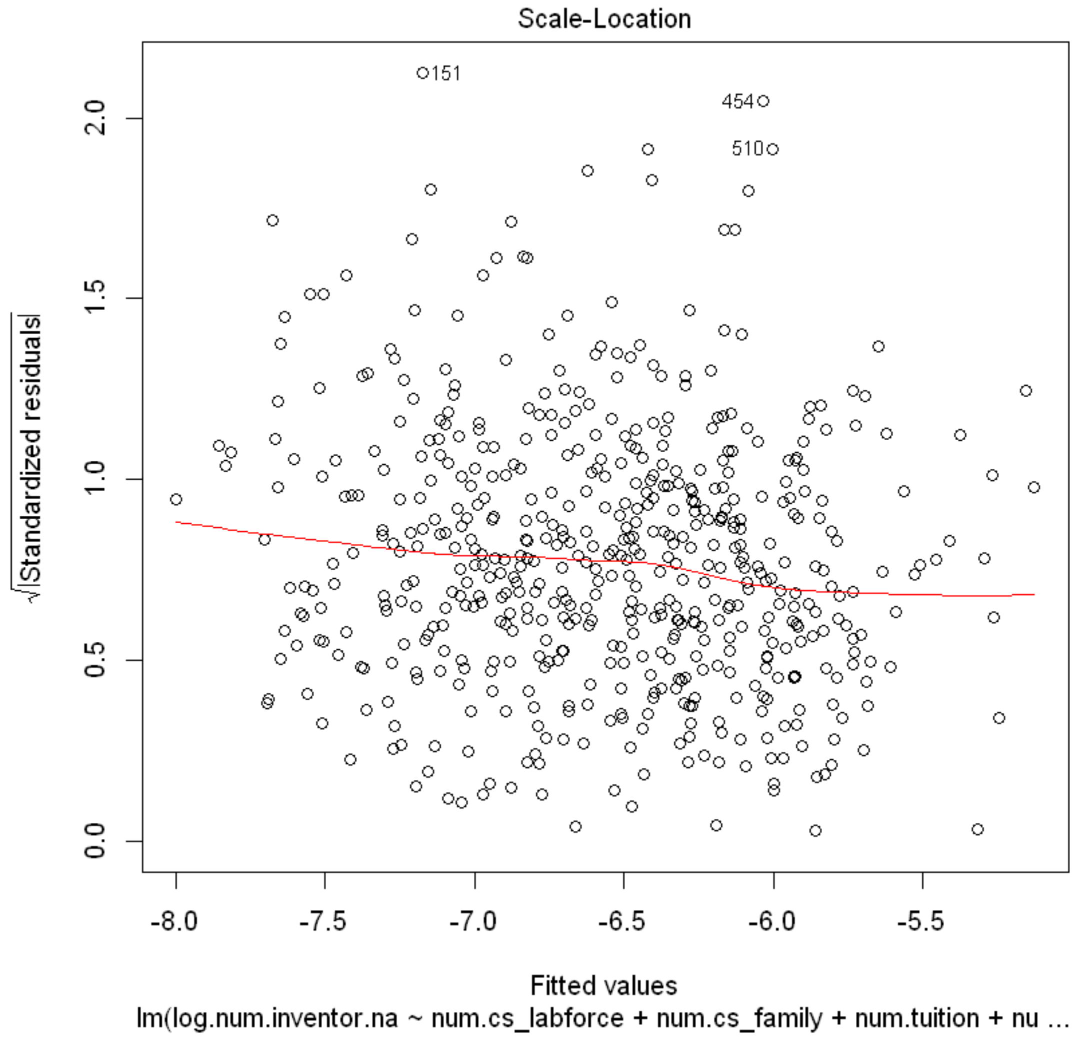


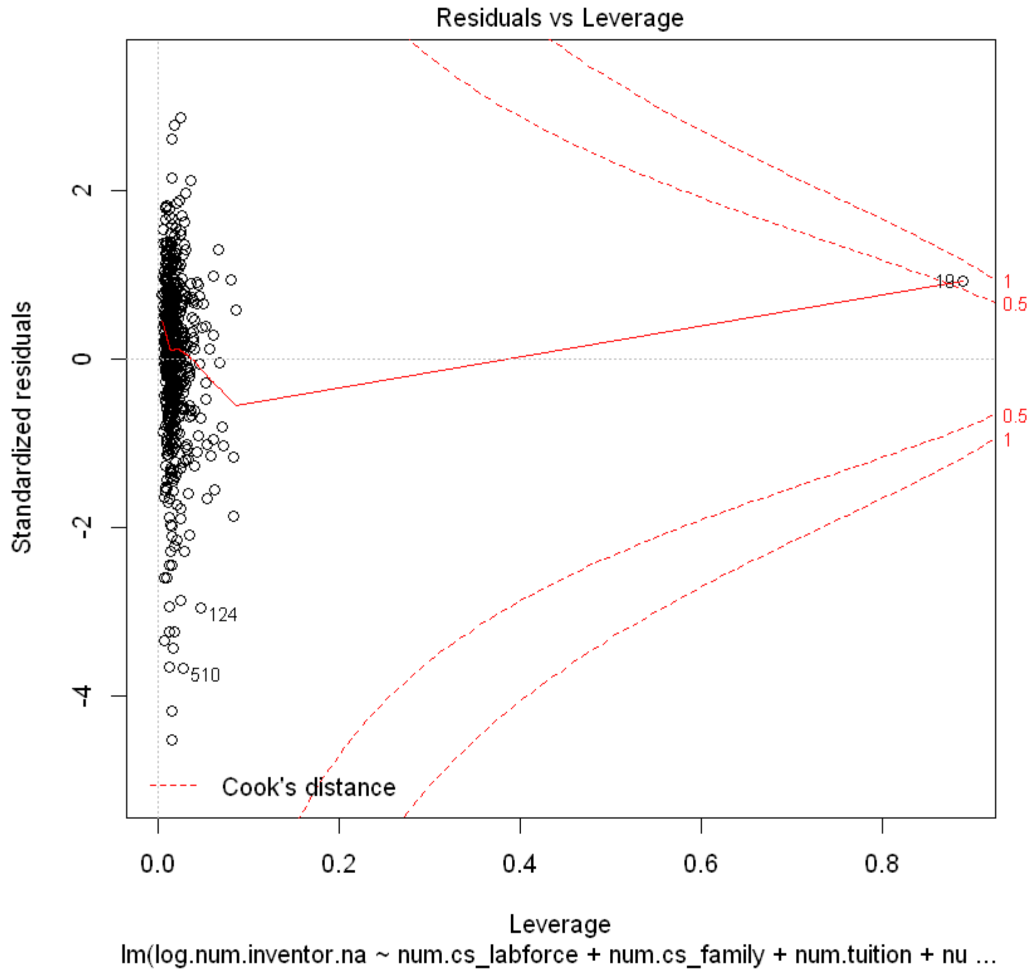
```
Call:
glm(formula = log.num.inventor.na ~ num.cs_labforce + num.cs_family + 
    num.tuition + num.cs_married + num.inc_share_1perc + num.inc_shar_1perc2 + 
    num.gini + num.hhinc00 + num.scap + num.par_stateabbrv)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.01325  -0.22754   0.04612   0.28496   1.26734  

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -3.201e+00  7.060e-01  -4.534 7.13e-06 ***
num.cs_labforce      2.711e-01  5.010e-01   0.541 0.588618    
num.cs_family       -6.352e+00  7.653e-01  -8.300 8.32e-16 ***
num.tuition          3.983e-06  5.598e-06   0.711 0.477086    
num.cs_married      -6.128e+00  7.394e-01  -8.287 9.13e-16 ***
num.inc_share_1perc  2.576e-02  1.642e-02   1.569 0.117313    
num.inc_shar_1perc2 -2.969e-04  2.994e-04  -0.992 0.321741    
num.gini            -1.864e+00  6.535e-01  -2.852 0.004511 ** 
num.hhinc00          5.043e-05  4.919e-06  10.252  < 2e-16 ***
num.scap             1.059e-01  2.501e-02   4.235 2.68e-05 ***
num.par_stateabbrv   4.781e-03  1.375e-03   3.477 0.000547 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2018428)

    Null deviance: 274.69  on 553  degrees of freedom
Residual deviance: 109.60  on 543  degrees of freedom
  (187 observations deleted due to missingness)
AIC: 698.53

Number of Fisher Scoring iterations: 2
```

Regression Diagnostics






Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
151
454
510
-8.0
-7.5
-7.0
-6.5
-6.0
-5.5
Fitted values
lm(log.num.inventor.na ~ num.cs_labforce + num.cs_family + num.tuition + nu ...


Residuals vs Leverage
Standardized residuals
-4
-2
0
2
19
124
510
1
0.5
0.5
1
Cook's distance
0.0
0.2
0.4
0.6
0.8
Leverage
lm(log.num.inventor.na ~ num.cs_labforce + num.cs_family + num.tuition + nu ...

NULL

Packages

- dplyr
- readstata13
- usmap
- ggplot2
- regtools
- tidyverse

Attachment Folder Contents

- Working Files
- Original Data Sets
- Text File for R Codes
- HTML Files
- Results Folder

Contact Information

Megha S Patel
732-762-8181
patel.megha@bls.gov
https://github.com/msp156
https://www.linkedin.com/in/megha-patel-01a376a8