IE 590 Predictive Modeling
Lab 4 Assignment
Chen Yung-Ching

◆ Problem 1

Table 1. The N/A ratio and other information for each dataframe

	df.names	na.ratio	df.row	df.col
1	creative.df	0.000000000	3146	20
2	education.df	0.079476151	3662	48
3	population.df	0.009020261	3275	133
4	poverty.df	0.088889094	3194	34
5	unemployment.df	0.001685261	3275	52

Firstly, I computed the N/A ratio for each dataframe. The definition of N/A ration here is simply total number of N/A divided by total cells that a dataframe has. From the result in Table 1. we see that `creative.df` is the most complete one and it even doesn't contain any N/A value. On the other hand, 8.89% of data in `poverty.df` is N/A, which is the highest among these five dataframes. Thus, we take a deeper look at this dataset, and we can find out that the data in columns `POV05_2016, CI90LB05_2016, CI90UB05_2016, PCTPOV05_2016, CI90LB05P_2016, CI90UB05P_2016` is very scarce. That is the reason why N/A ratio in this dataset turns out to be the highest.

By just looking at data sets themselves, I would somehow trust these data because in each dataframe there are a lot of details that seems reasonble.

◆ Problem 2

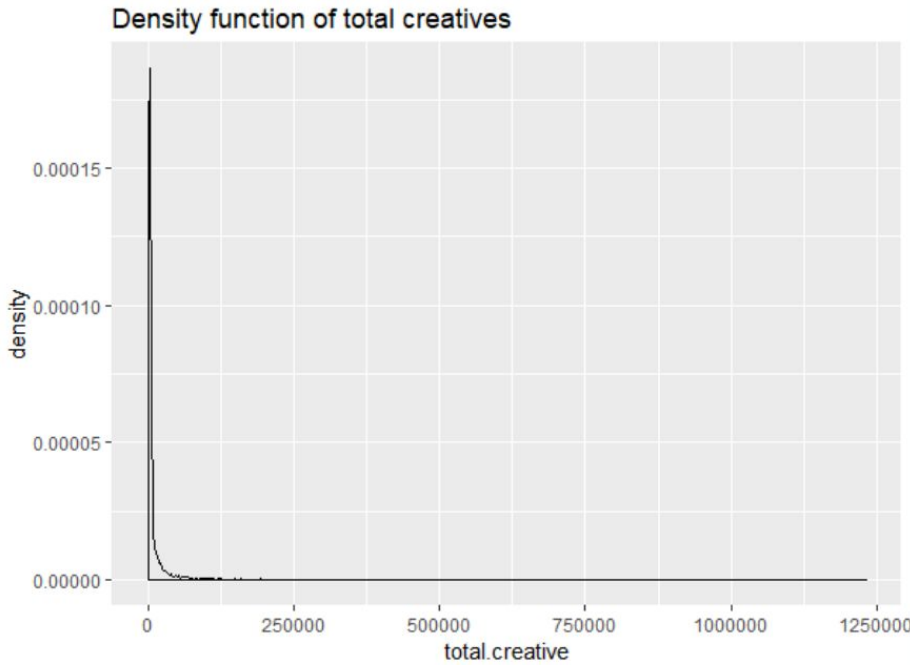

Figure 1. The density function of total creatives

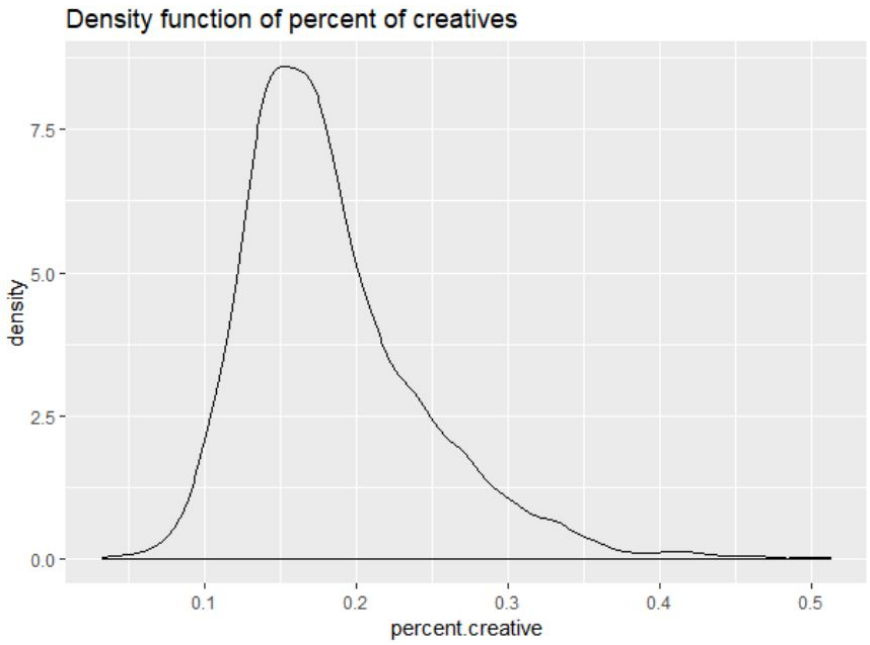

Figure 2. The density function of percent of creatives

I might not use a fraction of creatives versus total number of creatives in a model because fraction of creatives is computed by dividing total number of creatives by total number of employment. In other word, the value of ratio is actually based on total number of creatives and total number of employment. Solely using number of creatives as predictor to predict fraction of creatives is not a reasonable way.

◆ Problem 3

In this part, I merged different columns in `creative.df`, `education.df`, `population.df`, `poverty.df`, `unemployment.df` into another dataframe that may be useful to predict total number of creatives. To be more specific, there are totally 3133 rows and 39 columns in the new merged dataframe. To make it easier to read and understand, I also change some columns into other more reasonable and unified names. In Table.2 there are columns in new dataframe.

Table 2. The columns in new merged dataframe

```
# FIPS: state-county FIPS code
# state: state name
# state.abr: state abbreviation
# country: country name
# metro: metro area or not in 2003
# total.emp: total num of civilian employed, 2007-11
# total.creative: total num of creative employed, 2007-11
# percent.creative: percent of creative employed , 2007-11
# less.than.highschool: less than a high school diploma, 2012-16
# highschool.only: high school diploma only, 2012-16
# college: college or associate's degree, 2012-16
# bachelor.higher: bachelor's degree or higher, 2012-16
# percent.less.than.highschool: percent of adults with less than a high
school diploma, 2012-16
# percent.highschool.only: percent of adults with a high school diploma
only, 2012-16
# percent.college: percent of adults completing some college or
associate's degree, 2012-16
# percent.bachelor.higher: percent of adults with a bachelor's degree or
higher, 2012-16
# pop.2010: resident total population estimate, 2010
# pop.2011: resident total population estimate, 2011
# birth.2010: number of birth, 2010
# birth.2011: number of birth, 2011
# international.mig.2010: net international migration, 2010
# international.mig.2011: net international migration, 2011
# domestic.mig.2010: net domestic migration, 2010
# domestic.mig.2011: net domestic migration, 2011
# all.pov.2016: people of all ages in poverty, 2016
# percent.all.pov.2016: percent of people of all ages in poverty, 2016
# 0.17.pov.2016: people age 0-17 in poverty, 2016
# percent.0.17.pov.2016: percent of people age 0-17 in poverty, 2016
# median.income.2016: estimate of median household income 2016
# employed.2007: number employed annual average, 2007
# unemployed.2007: number unemployed annual average, 2007
# employed.2008: number employed annual average, 2008
# unemployed.2008: number unemployed annual average, 2008
# employed.2009: number employed annual average, 2009
# unemployed.2009: number unemployed annual average, 2009
# employed.2010: number employed annual average, 2010
# unemployed.2010: number unemployed annual average, 2010
# employed.2011: number employed annual average, 2011
# unemployed.2011: number unemployed annual average, 2011
```

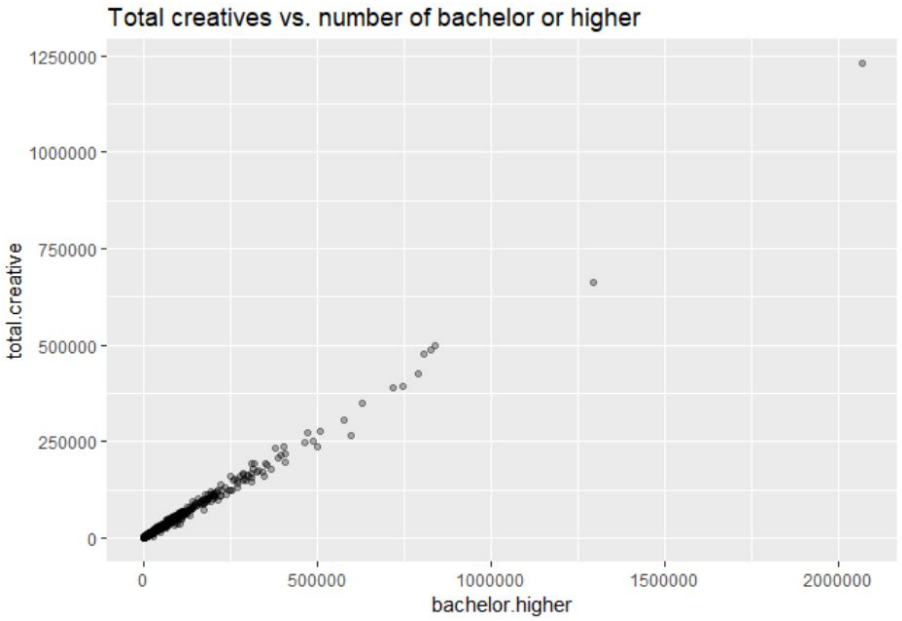

Figure 3. The scatter plot of total number of creatives versus total number of bachelor or higher degrees

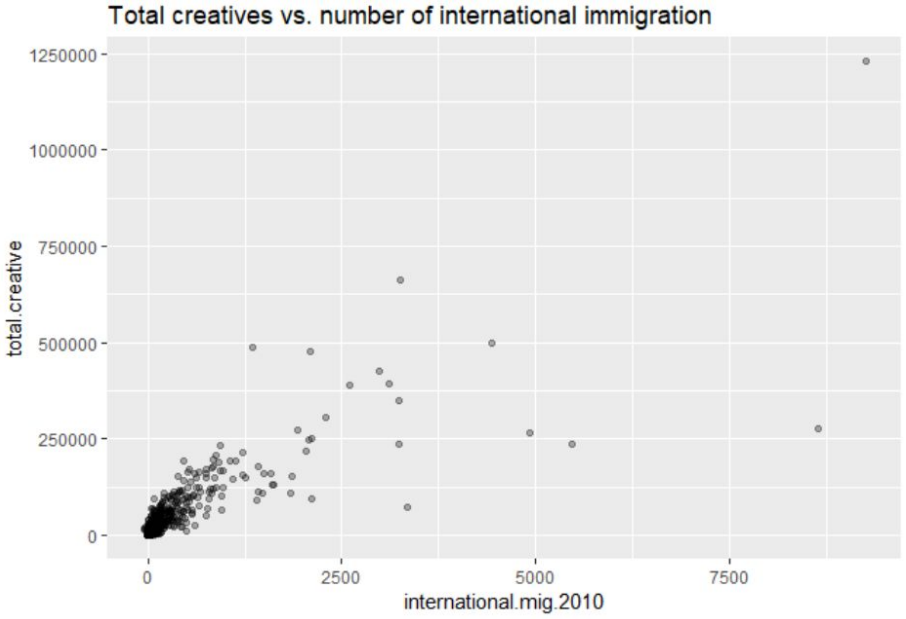

Figure 4. The scatter plot of total number of creatives versus total number of internation immigration

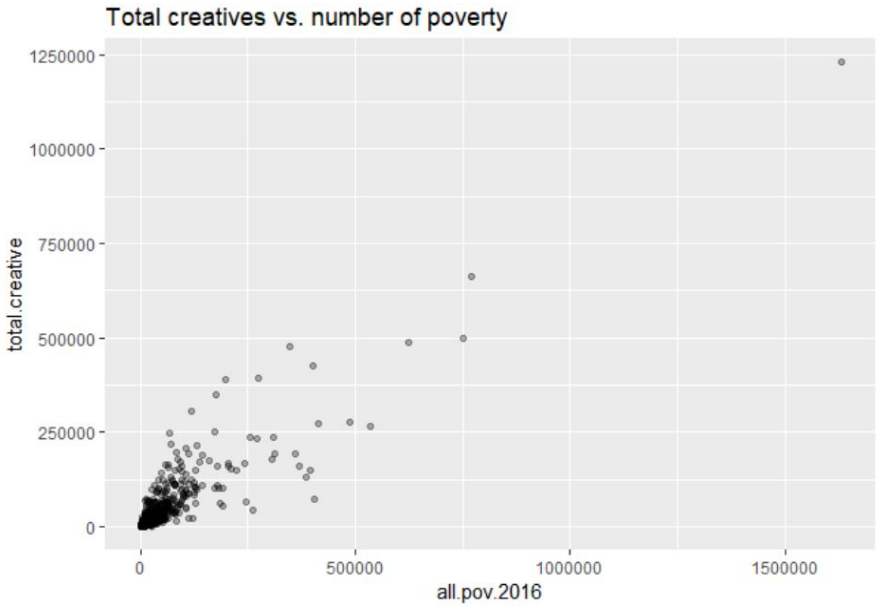

Figure 5. The scatter plot of total number of creatives versus total number of poverty

Then I explore the relationship between some potential predictors and response variable by interpreting the plots. From the plots, we may have the following three hypotheses.

(1) The relationship between the number of adults with a bachelor degree or higher 2012-2016 and total number of creatives is positive.

(2) The relationship between net international migration 2010 and total number of creatives is positive.

(3) The relationship between estimate of people of all ages in poverty 2016 and total number of creatives is negative.

◆ Problem 4

Before building linear regression models, I transform `metro` into factor type and I also split the 85% of data into training set and the rest into test set. Then I tried to build several simple linear regression models first to explore the relationships between different predictors and response variable `total.creative`, and also check whether our previous hypotheses are valid or not.

Table 3. The adjusted R-squared in different simple linear regression models

Model Name	Simple Linear Regression Model	Adjusted R-Squared	RMSE on Training set	RMSE on Test set
simple.1	total.creative ~ total.emp	0.9703	7630.828	6566.775
simple.2	total.creative ~ bachelor.higher	0.9941	3415.406	3860.82
simple.3	total.creative ~ international.mig.2010	0.7277	23108.895	21223.374

simple.4	total.creative ~ all.pov.2016	0.8321	18142.538	15732.701
simple.5	otal.creative ~ median.income.2016	0.1047	41900.631	28383.071

After we have initial idea of some variables, we may forward to building multiple regression models. The first model I try to build includes the following 20 columns which are based on the preliminary analysis in the previous part and also some of my personal assumptions that they may be good predictors or worth being included to answer some questions.

Table 4. The predictors in multiple regression model m1 and the corresponding assumptions

Predictors	Assumption & Possible Questions
`metro`	Metro area may have more creatives.
`total.emp`	Higher employment may have higher creatives.
`less.than.highschool`, `highschool.only` , `college`, `bachelor.higher`	The education level may influnece creatives. E.g. higher bachelor may result in higher creatives.
`birth.2010`, `birth.2011`	Whether # of birth have any relationship with creatives
`international.mig.2010`, `international.mig.2011`, `domestic.mig.2010`, `domestic.mig.2011`	The demographic change may also inflect one area's creatives. E.g. higher international immigration may mean higher creatives
`all.pov.2016`, `` `0.17.pov.2016` ``	Higher poverty may have less creatives
`median.income.2016`	Higher income may have higher creatives
`unemployed.2007`, `unemployed.2008`, `unemployed.2009`, `unemployed.2010`, `unemployed.2011`	Higher umemployed may have lower creatives

Table 5. The summary result of multiple regression model m1

```
Residuals:
     Min       1Q   Median       3Q      Max
-19263.9   -294.8    -67.8    296.0  24824.1

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             9.436e+01  1.586e+02   0.595  0.55198
```

```
metro1                    1.828e+02  8.474e+01   2.157  0.03107 *
total.emp                 2.501e-01  6.232e-03  40.128  < 2e-16 ***
less.than.highschool     -2.423e-02  8.094e-03  -2.993  0.00279 **
highschool.only          -1.966e-01  5.810e-03 -33.845  < 2e-16 ***
college                  -6.477e-02  6.742e-03  -9.607  < 2e-16 ***
bachelor.higher           3.082e-01  5.481e-03  56.226  < 2e-16 ***
birth.2010               -2.069e+01  1.316e+00 -15.718  < 2e-16 ***
birth.2011                4.723e+00  3.643e-01  12.964  < 2e-16 ***
international.mig.2010   -3.632e+00  1.422e+00  -2.554  0.01070 *
international.mig.2011    4.271e-01  3.377e-01   1.265  0.20613
domestic.mig.2010         1.357e-01  1.239e-01   1.095  0.27357
domestic.mig.2011         1.890e-01  3.890e-02   4.859 1.25e-06 ***
all.pov.2016             -6.278e-02  1.020e-02  -6.152 8.91e-10 ***
`0.17.pov.2016`           4.142e-02  2.953e-02   1.402  0.16091
median.income.2016        3.787e-04  3.341e-03   0.113  0.90976
unemployed.2007          -5.184e-01  5.359e-02  -9.674  < 2e-16 ***
unemployed.2008           3.619e-02  7.844e-02   0.461  0.64456
unemployed.2009           2.310e-01  7.178e-02   3.219  0.00130 **
unemployed.2010           9.046e-01  1.211e-01   7.470 1.11e-13 ***
unemployed.2011          -9.057e-01  8.233e-02 -11.001  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1664 on 2485 degrees of freedom
Multiple R-squared:  0.9986,  Adjusted R-squared:  0.9986
F-statistic: 8.864e+04 on 20 and 2485 DF,  p-value: < 2.2e-16
```

From Table 5. we may say that `total.emp` may be one of important predictors. Since now we want to predict the number of creatives, this would be a useful information because without any other information, we may assume that the higher the number of total employment, the higher the total creatives the area may have. Other predictors related to education like `less.than.highschool`, `highschool.only`, `college`, `bachelor.higher` also seem to be important factors in this model. This is reasonable because usually creative jobs require people to have higher education background. Thus these education columns are helpful information to predict number of creatives. Lastly, the number of poverty in the area is also a good predictor. If the number of poverty is high in one area, then its number of creatives tends to be low.

On the other hand, `metro` is not a significant factor, so we may say whether an area is metropolitan or not is not an important information we should consider. Another insignificant independent variable is `median.income.2016`. At first we may assume that creative jobs may have higher income, thus the median income may

be a good predictor to predict number of creatives. However, from the model we built we see that the this assumption isn't too plausible. That is the median of income of an area may not be a good predictor to predict total number of creatives.

Based on the first model, let's build another one using stepwise method to do variable selection and see whether our interpretation is true or wrong. From the Table 6. we see that `metro` and `median.income.2016` are not chosen, indicating that they really may not be the important predictors in this case.

Next, let's see how fit and effective our models are. From Table 7. we can see that the adjusted R-squared on training set is pretty high. To see whether they are overfitting, I compare the RMSE on both training set and test set. Based on the computation results, the RMSE are similar both on training set and test set, meaning that the models can be generalized well to unseen data.

Table 6. The summary result of multiple regression model m2.stepwise

```
Residuals:
     Min       1Q   Median       3Q      Max
-19081.7   -290.1    -68.1    297.5  24520.4

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            108.760097  42.761588   2.543 0.011038 *
metro1                 183.604119  78.786264   2.330 0.019864 *
total.emp                0.249527   0.005771  43.238  < 2e-16 ***
less.than.highschool    -0.020708   0.007601  -2.724 0.006487 **
highschool.only         -0.195436   0.005601 -34.891  < 2e-16 ***
college                 -0.065243   0.006392 -10.208  < 2e-16 ***
bachelor.higher          0.310356   0.005256  59.043  < 2e-16 ***
birth.2010             -20.421112   1.241562 -16.448  < 2e-16 ***
birth.2011               4.617614   0.341094  13.538  < 2e-16 ***
international.mig.2010  -1.921650   0.295303  -6.507 9.21e-11 ***
domestic.mig.2011        0.219954   0.030954   7.106 1.56e-12 ***
all.pov.2016            -0.065073   0.009847  -6.609 4.73e-11 ***
`0.17.pov.2016`          0.050876   0.028326   1.796 0.072600 .
unemployed.2007         -0.498881   0.045269 -11.020  < 2e-16 ***
unemployed.2009          0.214384   0.061400   3.492 0.000488 ***
unemployed.2010          0.940041   0.112169   8.381  < 2e-16 ***
unemployed.2011         -0.915588   0.072870 -12.565  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1664 on 2489 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9986
F-statistic: 1.109e+05 on 16 and 2489 DF,  p-value: < 2.2e-16
```

Table 7. The adjusted R-squared in different multple linear regression models

Model Name	Adjusted R-Squared	RMSE on Training set	RMSE on Test set
m1	0.9984	1657.143	1657.952
m2.stepwise	0.9984	1926.555	1926.555

◆ Problem 5

In this part, I will predict the total creatives in three specific countries, Los Alamos County, Fairfax County and Robeson County to show how the models handle corner cases. In Table 8. we can see that the predictions of creatives on Alamos County and Fairfax County are close to the true values. As for Robeson County, the model has an underestimated result.

Table 8. The prediction of creatives versus true value of creatives for Los Alamos County, Fairfax County and Robeson County

County	Prediction of Creatives	True Value of Creatives
Los Alamos County	4587.734	4555
Fairfax County	244588.485	247540
Robeson County	4556.460	6300

◆ Problem 6

Lastly, I use the following codes to plot the map of percent of creatives in US.

```
map.df = new.df[,c('FIPS', 'percent.creative')]
colnames(map.df) = c('fips', 'value')
plot_usmap(regions = 'county', data=map.df, values='value') +
  scale_fill_continuous(name = 'percent.creative', label =
scales::comma) +
  theme(legend.position = 'right') +
  labs(title = 'Percent of Creatives in US')
```

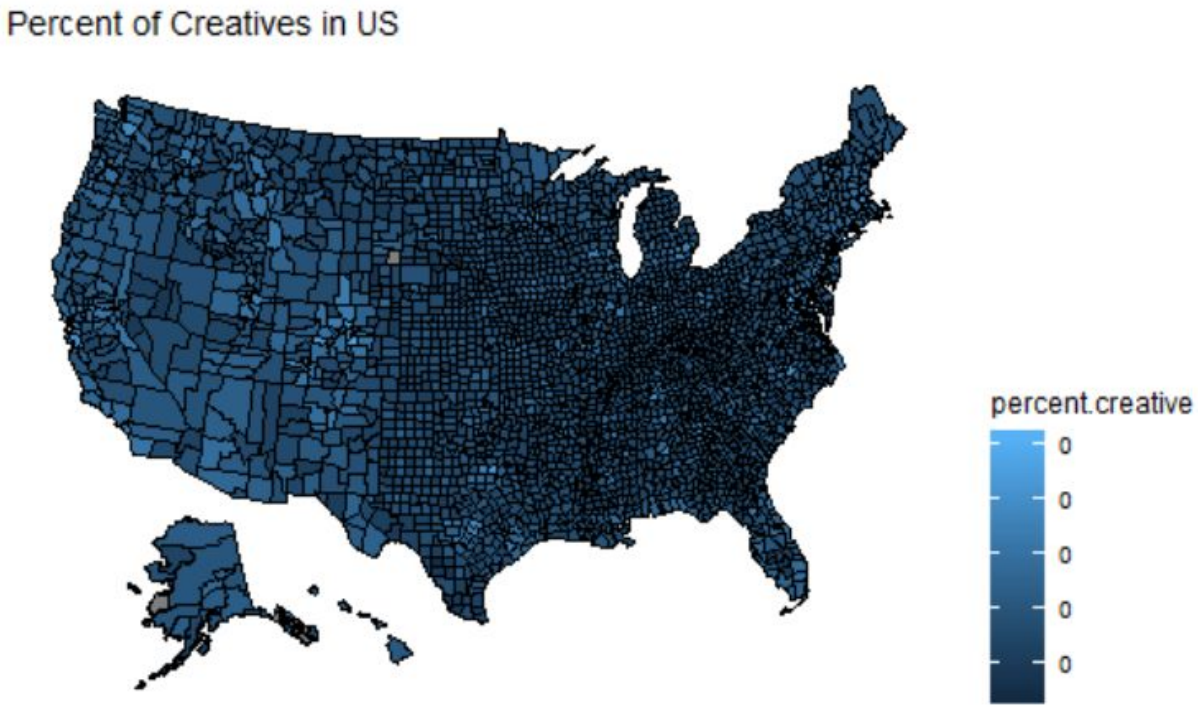

Figure 6. The heat map of percent of creatives in US