## Methods:

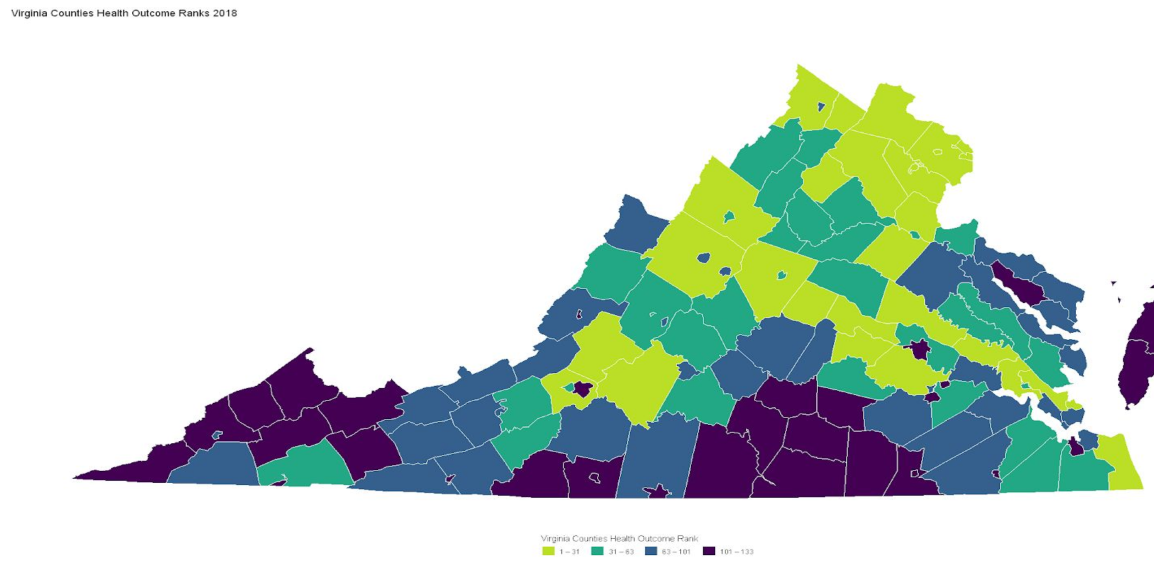

Virginia Counties Health Outcome Ranks 2018
Virginia Counties Health Outcome Rank
1 – 31
31 – 63
63 – 101
101 – 133

Figure 2: Distribution of Virginia’s Health Outcome Ranking

We generated the distribution map in Figure 2. It displays the distribution of Virginia’s health outcome ranking. The map is divided into four quartiles with less color intensity. Relatively less color intensity means better performance in the respective summary rankings. The Northern Virginia’s health condition ranks are very high compared to other counties.

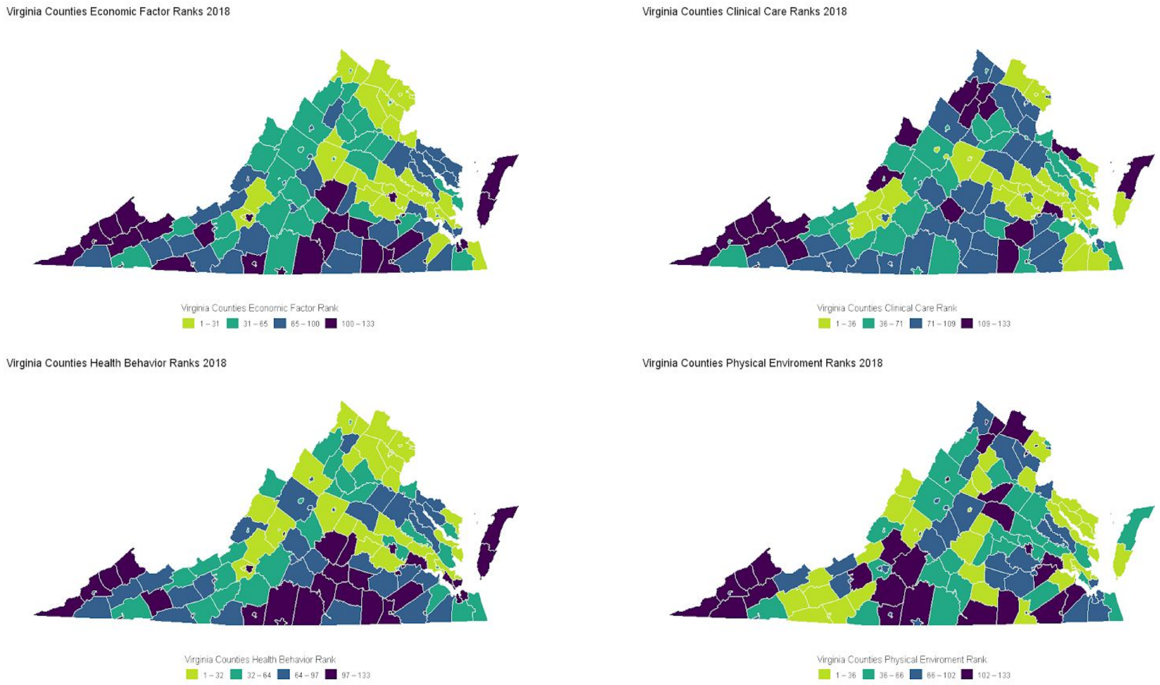

Virginia Counties Economic Factor Ranks 2018
Virginia Counties Economic Factor Rank
1 – 31
31 – 65
65 – 100
100 – 133
Virginia Counties Clinical Care Ranks 2018
Virginia Counties Clinical Care Rank
1 – 36
36 – 71
71 – 109
109 – 133
Virginia Counties Health Behavior Ranks 2018
Virginia Counties Health Behavior Rank
1 – 32
32 – 64
64 – 97
97 – 133
Virginia Counties Physical Enviroment Ranks 2018
Virginia Counties Physical Enviroment Rank
1 – 36
36 – 66
66 – 102
102 – 133

Figure 3: Distribution of Virginia’s Health Factor Ranking

The choropleth map above (Figure 3) shows the distribution of Virginia's four-health-factor ranking. Rankings are divided into four quartiles with shade in color intensity. The Northern Virginia area also ranks high in economical factor, health behavior, and clinical care. On the other hand, physical environment factor is not high in ranking.

We can see that the distribution of economical factors and health behavior ranking resembles more closely to health outcome distribution pattern. We also speculate that economics and health behavior factors probably have some influences over health outcome judging from the distribution map. We will now build a model to accurately predict the relationship.

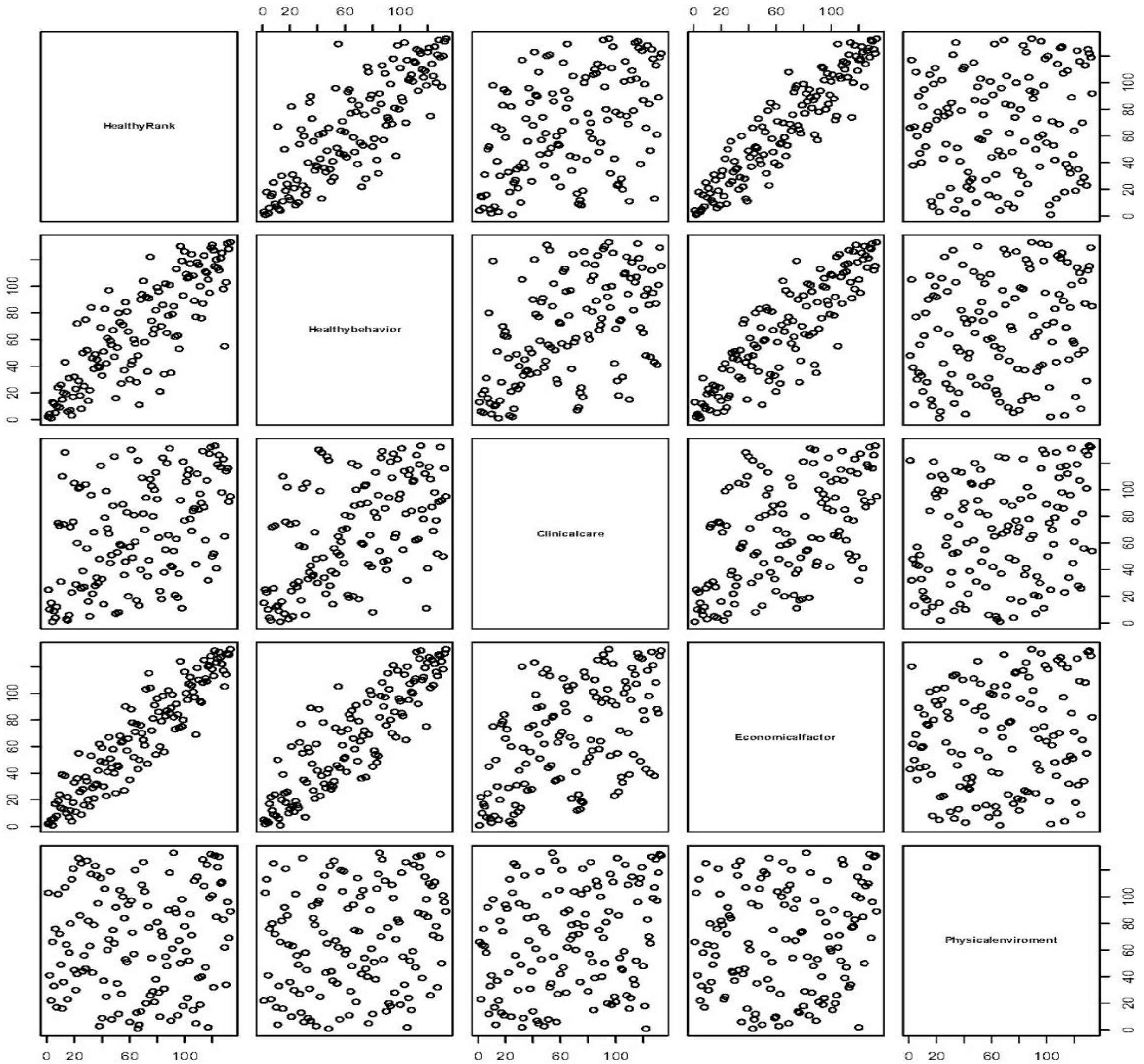
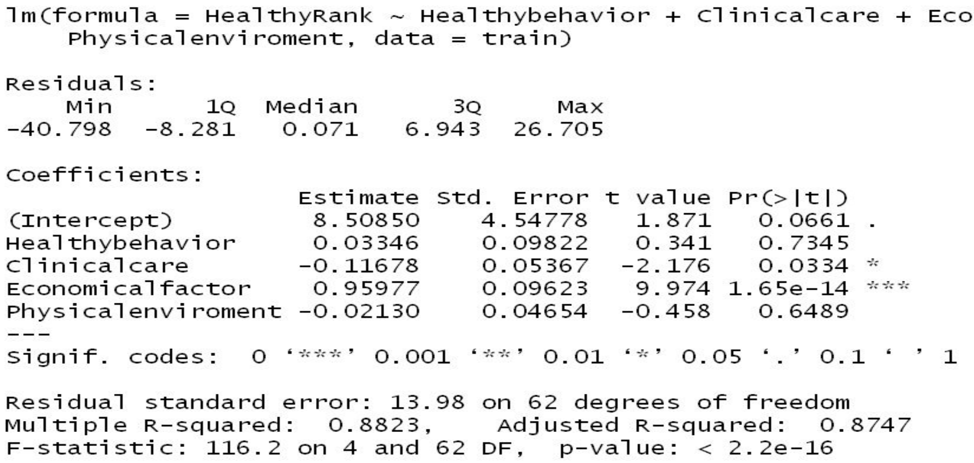


Figure 4: Scatter Matrix for Factors

A scatter matrix (Figure 4) is created in order to explore the potential relationships between variables. Economic factor is linearly correlated with health outcome. Health behavior is also positively correlated with health outcome. The other two factors are scattered and needed further model to predict the relationship.

**Model:**

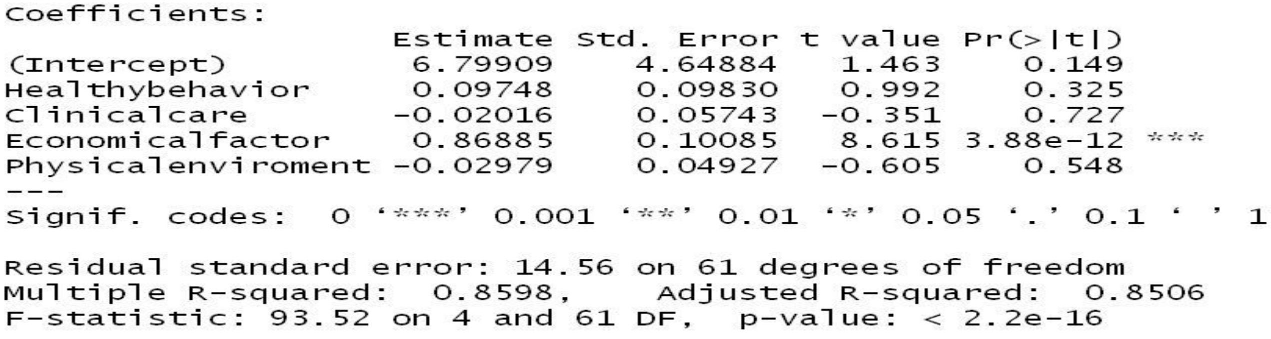
To avoid overfitting the model, we subset the data into two parts for training and testing (exactly 50%). We then put all four factors into the linear regression model. The result below (Figure 5) is the regression model on the training data. R square is 0.8823 which means that four factors can explain around 88% of the health outcome variabilities. Among these four factors, economic factor and clinical care's p-values are less than 0.05. This means that they are significant in predicting healthy rank.



```
lm(formula = HealthyRank ~ Healthybehavior + Clinicalcare + Economicalfactor + 
    Physicalenviroment, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.798  -8.281   0.071   6.943  26.705 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         8.50850    4.54778   1.871   0.0661 .  
Healthybehavior     0.03346    0.09822   0.341   0.7345    
Clinicalcare       -0.11678    0.05367  -2.176   0.0334 *  
Economicalfactor    0.95977    0.09623   9.974 1.65e-14 ***
Physicalenviroment -0.02130    0.04654  -0.458   0.6489    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.98 on 62 degrees of freedom
Multiple R-squared:  0.8823,	Adjusted R-squared:  0.8747 
F-statistic: 116.2 on 4 and 62 DF,  p-value: < 2.2e-16
```

Figure 5: Linear Regression Model for Healthy Rank

The screenshot below (Figure 6) is the regression model  with all predictors on the testing data. Compared to the training model, R square is reduced to 0.8598. We can see that the clinical care factor is not significant in the testing data (p-value > 0.05).



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         6.79909    4.64884   1.463    0.149    
Healthybehavior     0.09748    0.09830   0.992    0.325    
Clinicalcare       -0.02016    0.05743  -0.351    0.727    
Economicalfactor    0.86885    0.10085   8.615 3.88e-12 ***
Physicalenviroment -0.02979    0.04927  -0.605    0.548    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.56 on 61 degrees of freedom
Multiple R-squared:  0.8598,	Adjusted R-squared:  0.8506 
F-statistic: 93.52 on 4 and 61 DF,  p-value: < 2.2e-16
```

Figure 6: Regression Model on Testing Data

When we generated these models, we had some assumptions on the data. As we tested those assumptions in the diagnostic plots below (Figure 7).

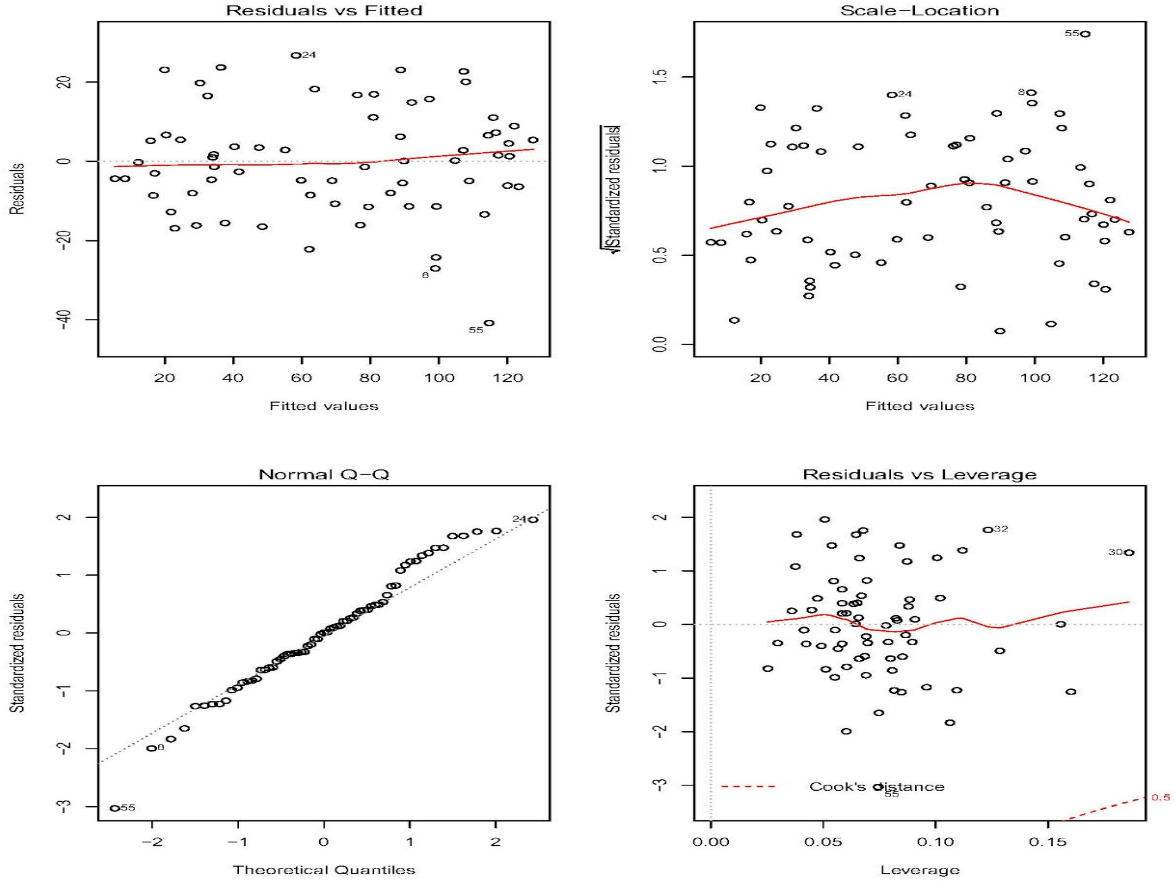
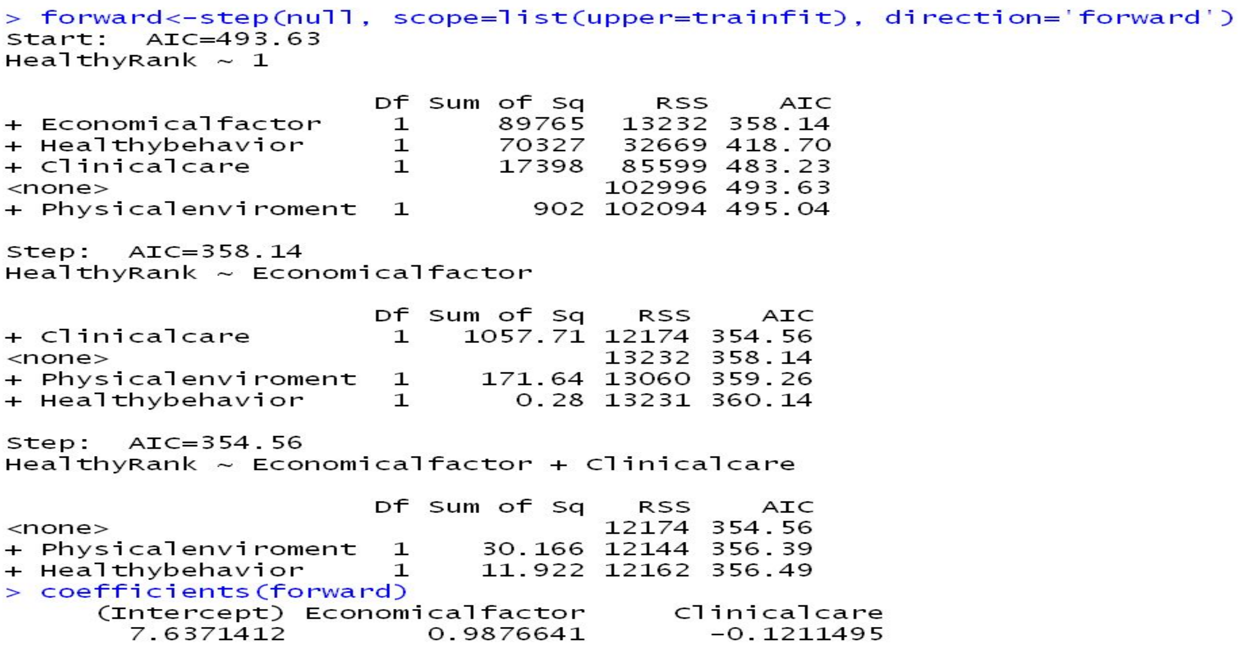


Figure 7: Diagnostic Plots

Residual versus Fitted line centers around line y=0. This means that our assumption about the model errors are independent. Model errors are also identically distributed in terms of mean value. For the Normal Q-Q plot, the data presents the outliers and a right tail which is located above the line. The residuals display normal distribution. While the Scale-Location plot, the curved line means the residual does not have the same variance. This indicates that our independent-identically-distributed-error assumption fails in terms of the variance. For the Residual-Leverage plot, there are no obvious high leverage points.

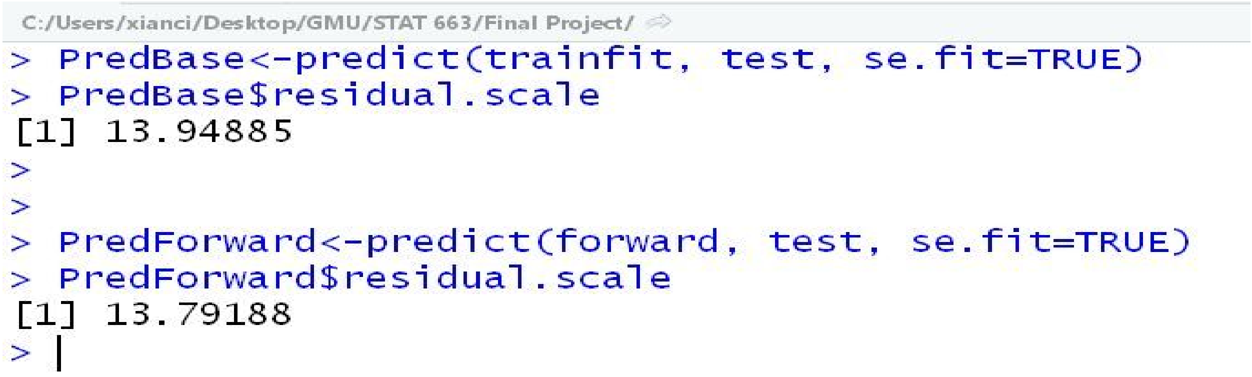
According to the diagnostics plots, we can summarize that the model fits the data. However, the testing model suggests that significance of some factor is varied. We decided to use the forward selection to figure out the most significant predictors.



```
> forward<-step(null, scope=list(upper=trainfit), direction='forward')
Start:  AIC=493.63
HealthyRank ~ 1

                      Df Sum of Sq    RSS    AIC
+ Economicalfactor     1     89765  13232 358.14
+ Healthybehavior      1     70327  32669 418.70
+ Clinicalcare         1     17398  85599 483.23
<none>                             102996 493.63
+ Physicalenviroment   1       902 102094 495.04

Step:  AIC=358.14
HealthyRank ~ Economicalfactor

                      Df Sum of Sq   RSS    AIC
+ Clinicalcare         1   1057.71 12174 354.56
<none>                             13232 358.14
+ Physicalenviroment   1    171.64 13060 359.26
+ Healthybehavior      1      0.28 13231 360.14

Step:  AIC=354.56
HealthyRank ~ Economicalfactor + Clinicalcare

                      Df Sum of Sq   RSS    AIC
<none>                             12174 354.56
+ Physicalenviroment   1    30.166 12144 356.39
+ Healthybehavior      1    11.922 12162 356.49
> coefficients(forward)
     (Intercept) Economicalfactor     Clinicalcare
       7.6371412        0.9876641       -0.1211495
```

Figure 8: Forward Selection Model

Forward selection model starts with no predictors in the model. It then iteratively adds the most significant predictors, and stops when the improvement is no longer statistically significant. Results above (Figure 8) shows the whole forward sub-selection process. When comparing these models, the smaller the AIC, the better the data fits. Eventually, we can see only two significant factors. A new model with only these two factors are created. We compare the original full model with the forward model and we can see that the forward model performs a little better than the full model (Figure 9).

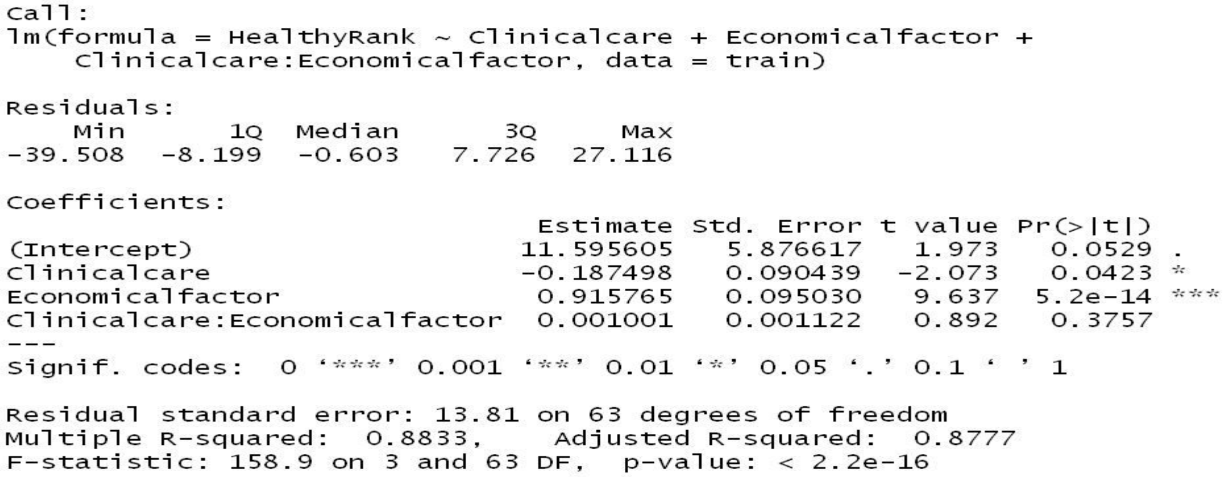


C:/Users/xianci/Desktop/GMU/STAT 663/Final Project/

```
> PredBase<-predict(trainfit, test, se.fit=TRUE)
> PredBase$residual.scale
[1] 13.94885
>
>
> PredForward<-predict(forward, test, se.fit=TRUE)
> PredForward$residual.scale
[1] 13.79188
> |
```

Figure 9: Residual Standard Deviations of The Models

Because of the process to explore data, we know that clinical care and economical factors are significant predictors to the health outcome. By observing the scatter matrix and diagnostic plots, we speculate that these two factors may interact with each other. Therefore, we added their interactions to the model (Figure 10). The R-square value is higher than the original full model.



```
Call:
lm(formula = HealthyRank ~ Clinicalcare + Economicalfactor +
    Clinicalcare:Economicalfactor, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-39.508  -8.199  -0.603   7.726  27.116

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  11.595605   5.876617   1.973   0.0529 .
Clinicalcare                 -0.187498   0.090439  -2.073   0.0423 *
Economicalfactor              0.915765   0.095030   9.637  5.2e-14 ***
Clinicalcare:Economicalfactor  0.001001   0.001122   0.892   0.3757
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.81 on 63 degrees of freedom
Multiple R-squared:  0.8833,    Adjusted R-squared:  0.8777
F-statistic: 158.9 on 3 and 63 DF,  p-value: < 2.2e-16
```

Figure 10: Forward Model with Interaction

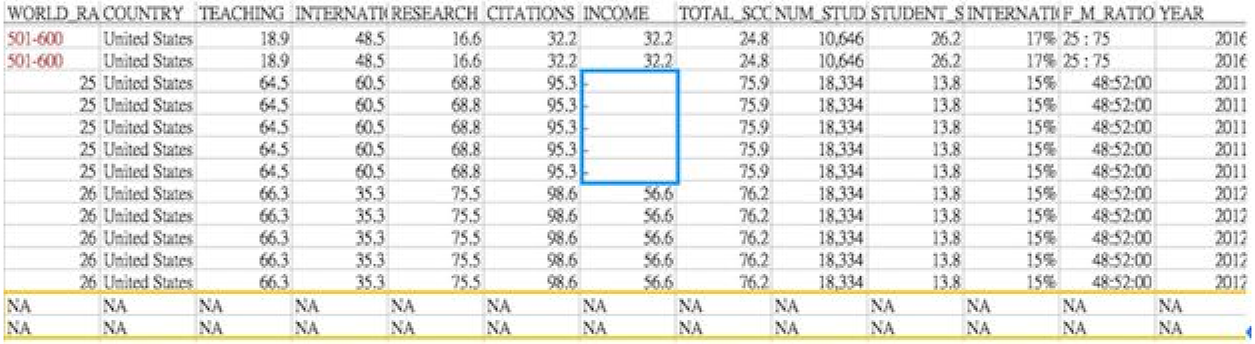
## 2. Data Science Programs in the United States

### Problem Description:

Because data science has become popular over the years, many organizations have started to hire data scientist as part of their organizations. Many potential students would want to know which higher education program could offer them the best outcome upon graduation. We want to explore factors that affect industry income in order to decide programs that have high quality and are innovative.

### Data Description:

The dataset is obtained from the list of datasets on Kaggle website [2]. The dataset was created from Times Education data set and data gathering on the websites by Kaggle author named Srihari Rao. The data set was originally created to merge all of the data science program offered in the United States for further exploration. The data set includes total 27 attributes which include school, state, city, program, type, department, world rank, teaching score, international score, research score, citation score, income, total score, number of students, student-to-staff ratio, international student percentage, female-to-male ratio, and year.



| WORLD_RA | COUNTRY | TEACHING | INTERNATI | RESEARCH | CITATIONS | INCOME | TOTAL_SCC | NUM_STUD | STUDENT_S | INTERNATI | F_M_RATIO | YEAR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 501-600 | United States | 18.9 | 48.5 | 16.6 | 32.2 | 32.2 | 24.8 | 10,646 | 26.2 | 17% | 25 : 75 | 2016 |
| 501-600 | United States | 18.9 | 48.5 | 16.6 | 32.2 | 32.2 | 24.8 | 10,646 | 26.2 | 17% | 25 : 75 | 2016 |
| 25 | United States | 64.5 | 60.5 | 68.8 | 95.3 | - | 75.9 | 18,334 | 13.8 | 15% | 48:52:00 | 2011 |
| 25 | United States | 64.5 | 60.5 | 68.8 | 95.3 | - | 75.9 | 18,334 | 13.8 | 15% | 48:52:00 | 2011 |
| 25 | United States | 64.5 | 60.5 | 68.8 | 95.3 | - | 75.9 | 18,334 | 13.8 | 15% | 48:52:00 | 2011 |
| 25 | United States | 64.5 | 60.5 | 68.8 | 95.3 | - | 75.9 | 18,334 | 13.8 | 15% | 48:52:00 | 2011 |
| 25 | United States | 64.5 | 60.5 | 68.8 | 95.3 | - | 75.9 | 18,334 | 13.8 | 15% | 48:52:00 | 2011 |
| 26 | United States | 66.3 | 35.3 | 75.5 | 98.6 | 56.6 | 76.2 | 18,334 | 13.8 | 15% | 48:52:00 | 2012 |
| 26 | United States | 66.3 | 35.3 | 75.5 | 98.6 | 56.6 | 76.2 | 18,334 | 13.8 | 15% | 48:52:00 | 2012 |
| 26 | United States | 66.3 | 35.3 | 75.5 | 98.6 | 56.6 | 76.2 | 18,334 | 13.8 | 15% | 48:52:00 | 2012 |
| 26 | United States | 66.3 | 35.3 | 75.5 | 98.6 | 56.6 | 76.2 | 18,334 | 13.8 | 15% | 48:52:00 | 2012 |
| 26 | United States | 66.3 | 35.3 | 75.5 | 98.6 | 56.6 | 76.2 | 18,334 | 13.8 | 15% | 48:52:00 | 2012 |
| NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |

Figure 11: Data Processing of the Dataset