MODELLING CANCER MORTALITY USING OLS REGRESSION

DA 546 PROJECT, JAN-APR, 2022

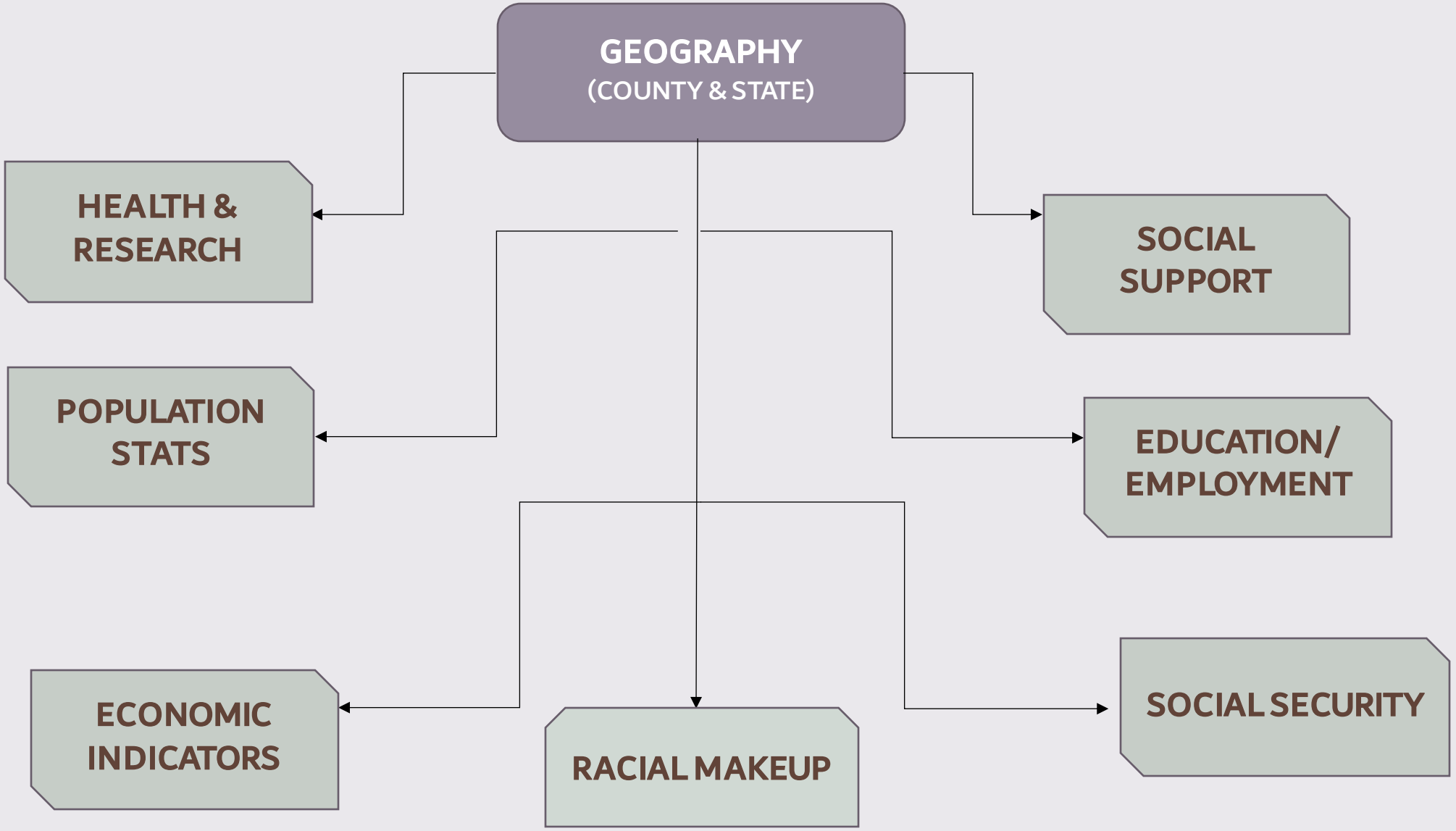
BUDDI KIRAN CHAITANYA
214161002
buddi.kiran@iitg.ac.in

INTRODUCTION:

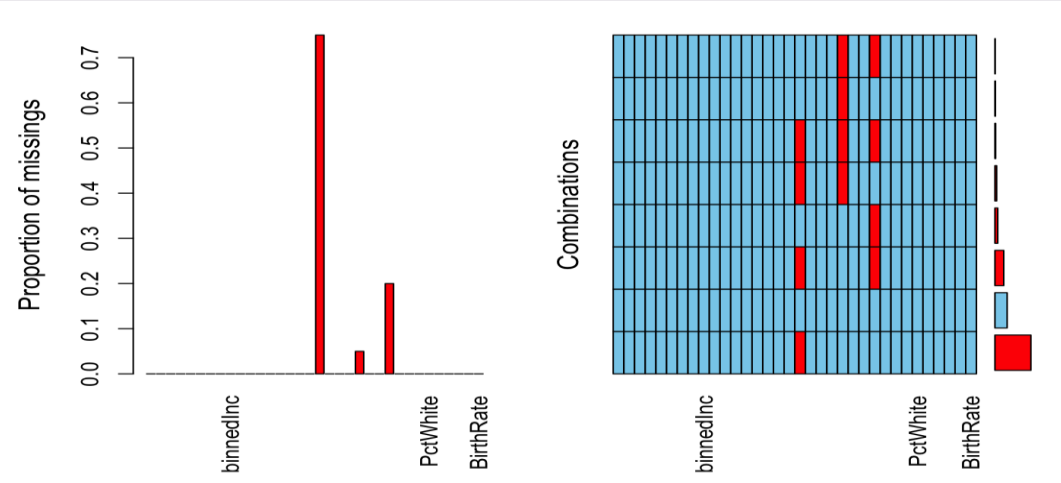
- -PROMLEM STATEMENT
- -SCOPE AND FEASIBILITY
- -WHY MORTALITY
- -ROADMAP & EXPECTATIONS

DATASET:

- -SOURCE
- -GENERAL DESCRIPTION OF THE VARIABLES
- -MISSING DATA & POSSIBLE ANAMOLIES
- -MODIFICATIONS


GEOGRAPHY
(COUNTY & STATE)
HEALTH & RESEARCH
SOCIAL SUPPORT
POPULATION STATS
EDUCATION/ EMPLOYMENT
ECONOMIC INDICATORS
RACIAL MAKEUP
SOCIAL SECURITY

MISSING DATA & MISC. :



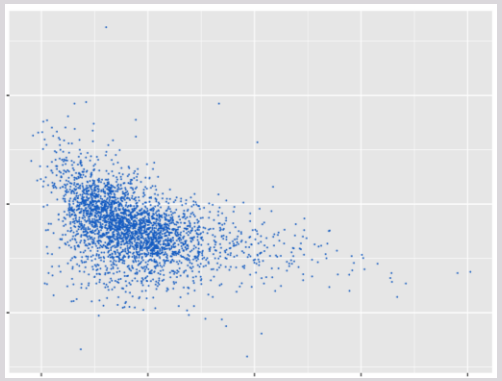
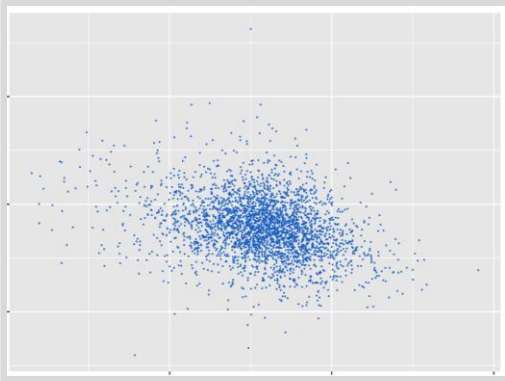
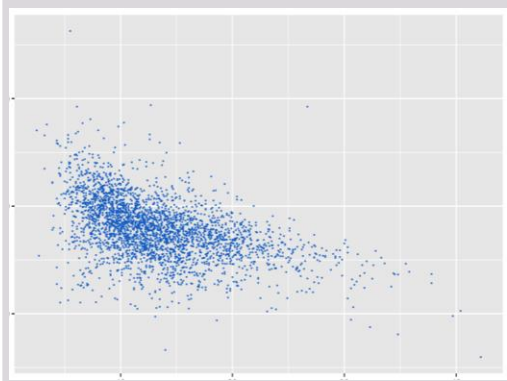
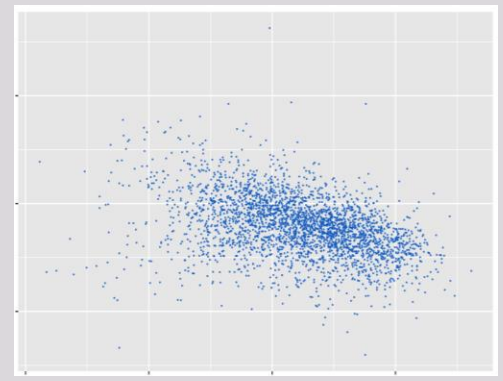
- MISSING DATA
- 3 ATTRIBUTES
- RECOURSE
- JUSTIFICATION

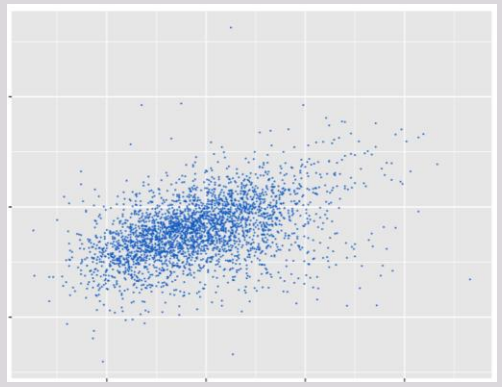
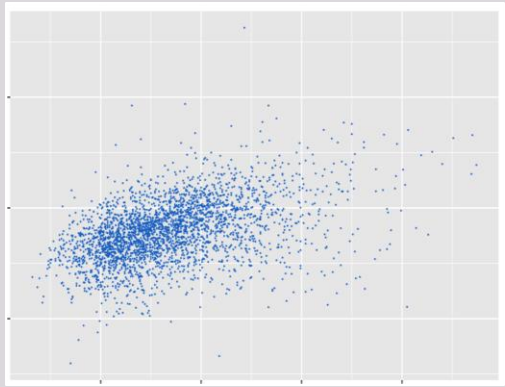
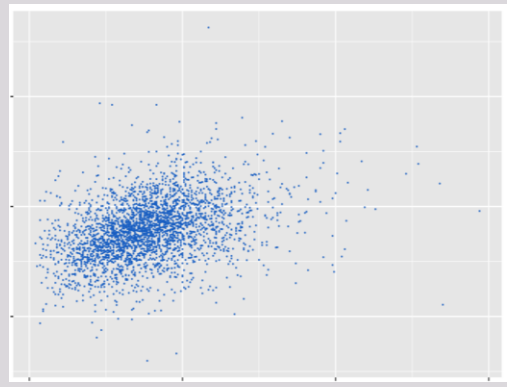
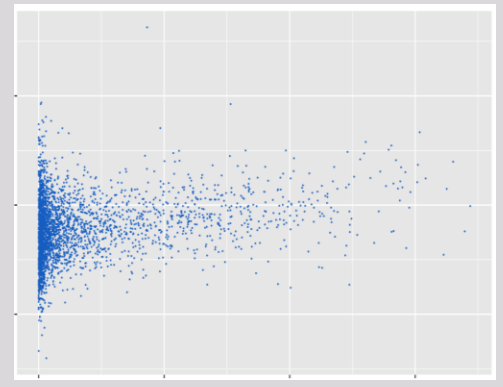
ANAMOLIES

HOUSEHOLD SIZE AND AGE

CORRECTIONS

SCATTER-PLOTS:

Med. Income	Pct. Marr Hlds.	Pct. >25. Bach	Pct. Pvt Cov
			

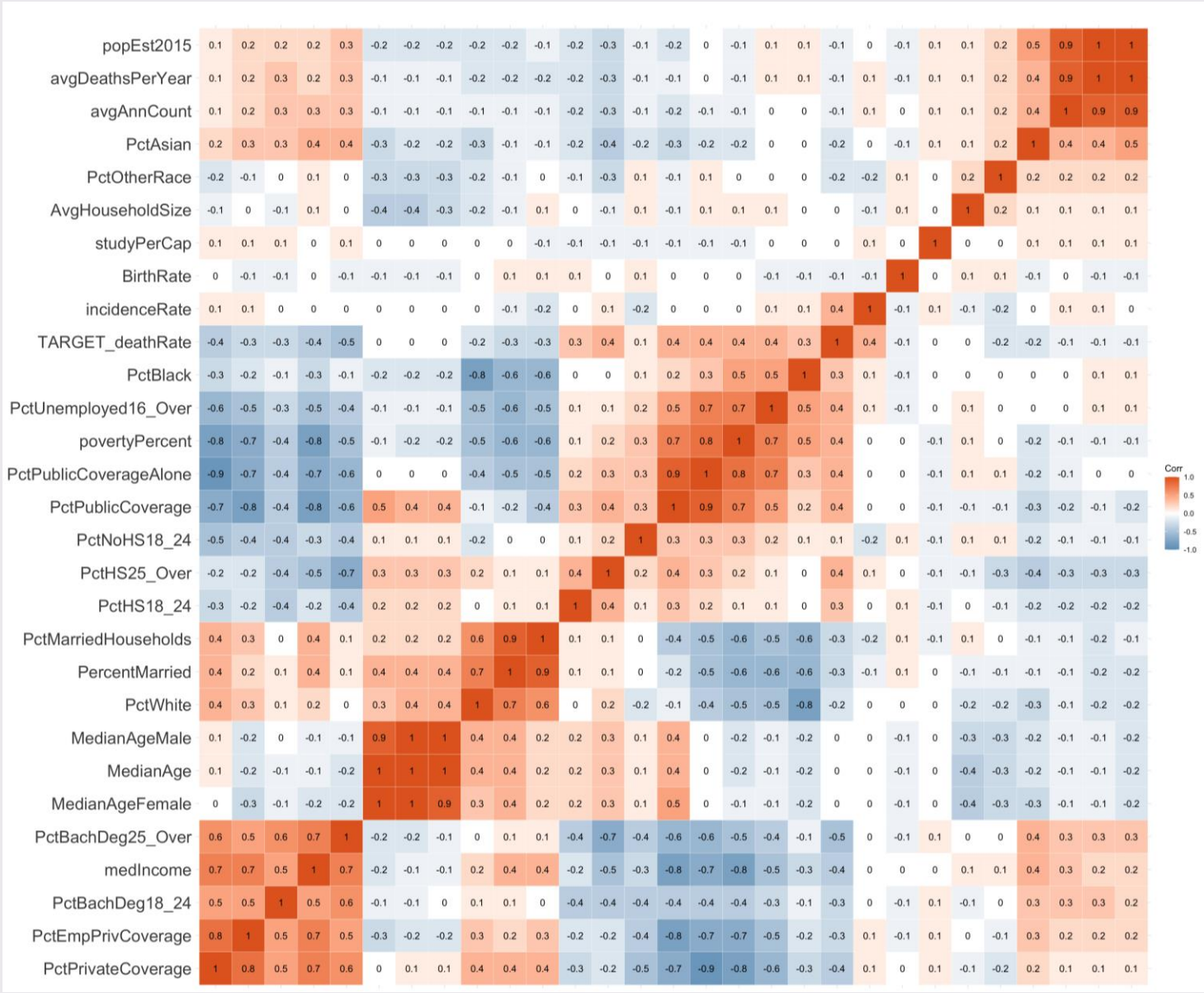
Pct. PBL Cv Alone	POVERTY PCT.	Pct. Unempl.	Pct. Black
			

HAND-PICKING FEATURES

MANUAL PUNING USING DATA CLUSTERS

- COVERAGE CLUSTER
- AGE CLUSTER
- MARRIED CLUSTER
- POVERTY CLUSTER
- POPULATION CLUSTER

HC-CORRELATION HEAT MAP:



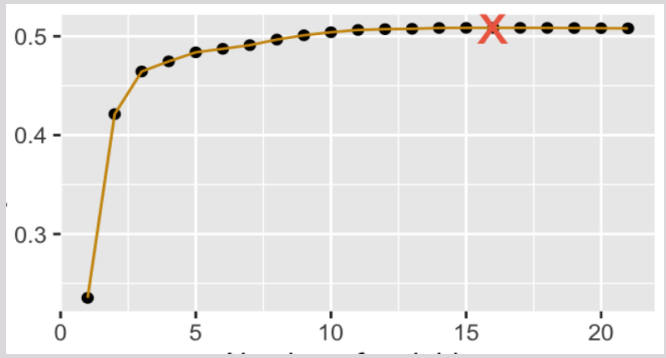
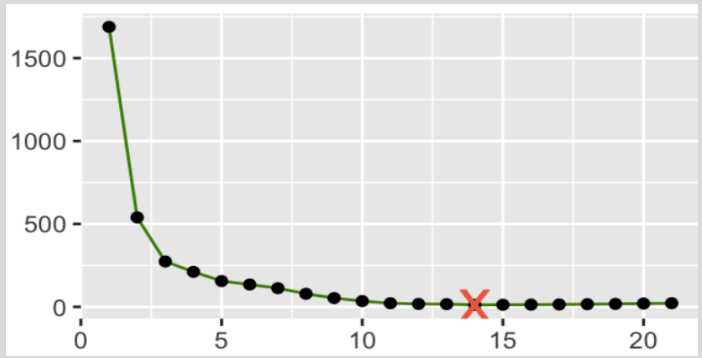
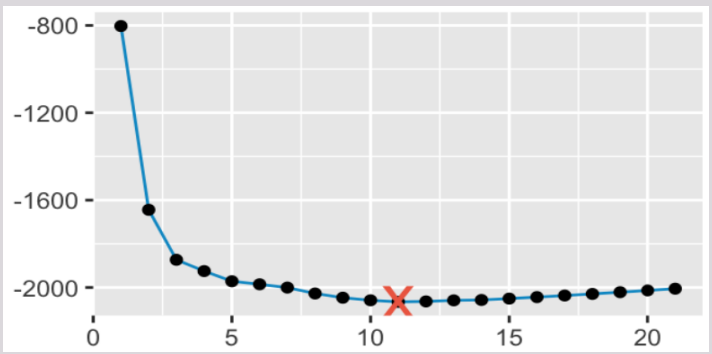
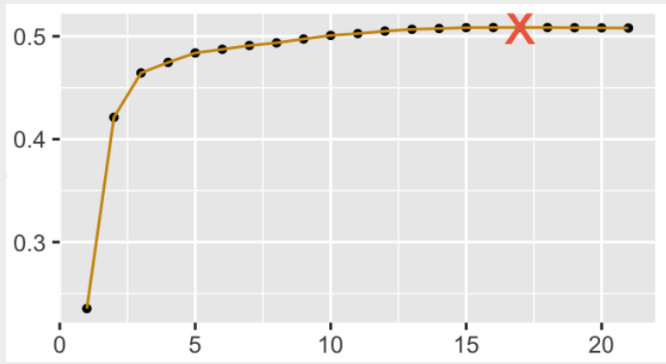
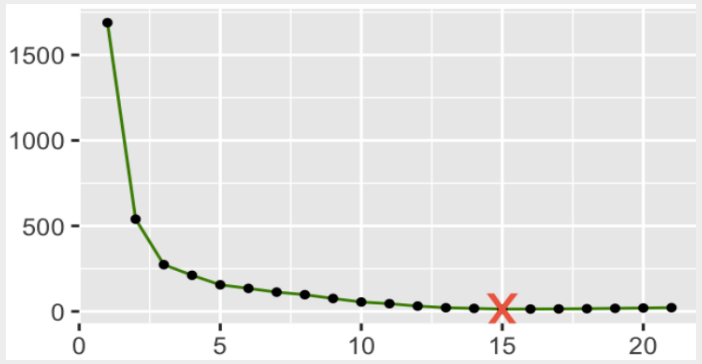
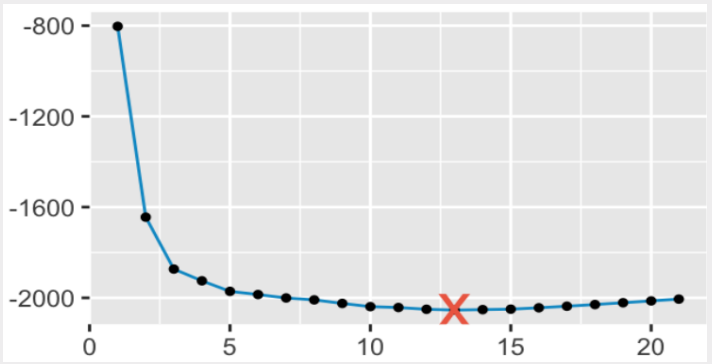
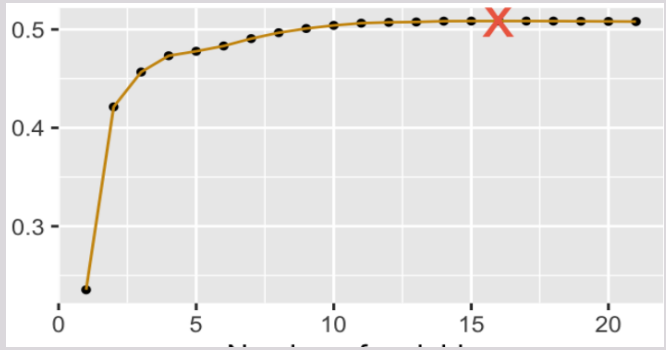
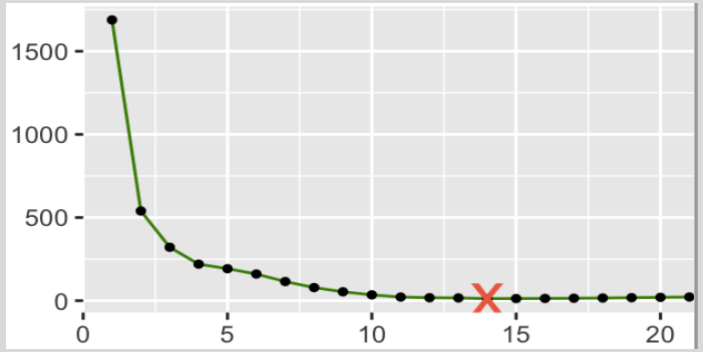
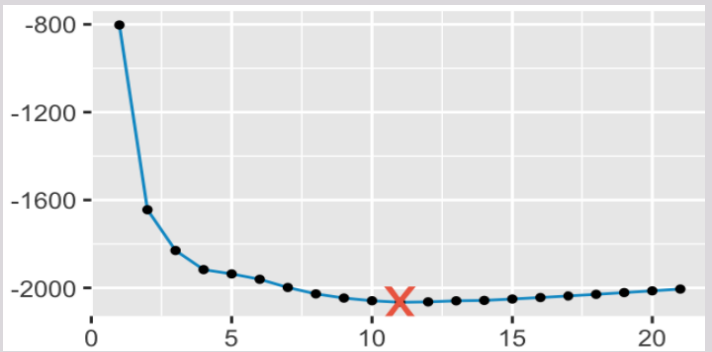
FULL FIT RESULTS WITH 21 VARIABLES:

```
## Residual standard error: 19.47 on 3025 degrees of freedom
## Multiple R-squared:  0.5114, Adjusted R-squared:  0.508
## F-statistic: 150.7 on 21 and 3025 DF,  p-value: < 2.2e-16
```

F-STATISTIC (>>1): SUFFICIENT (as n>>p) EVIDENCE AGAINST NUL HYPOTHESIS
Adj. R-SQUARED: 0.51

SUBSET SELECTION:

- BEST
- FORWARD
- BACKWARD
- VALIDATION SET APPROACH

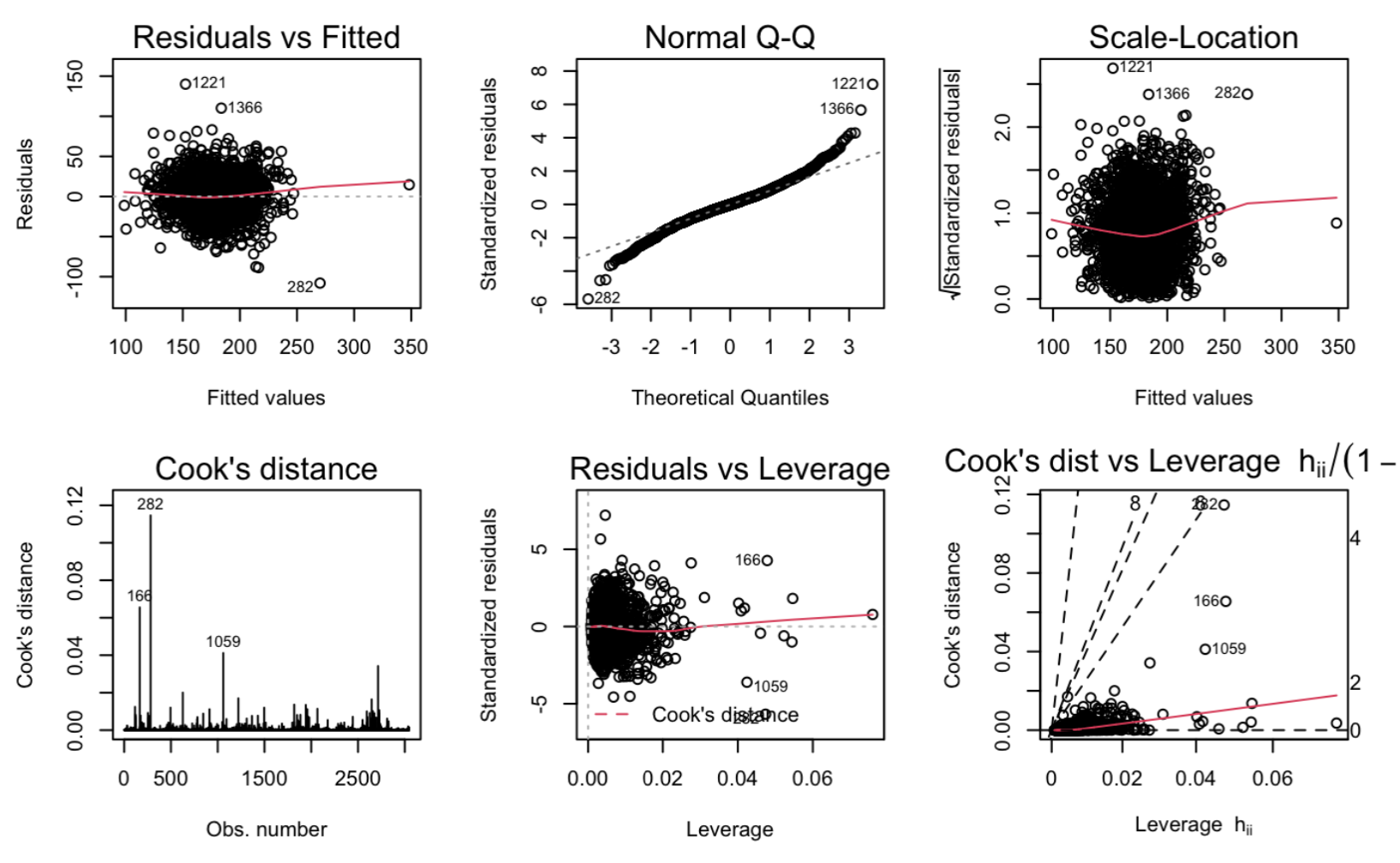
	ADJ R2	MALLOW'S Cp	BIC
BEST			
FORWARD			
BACKWARD			

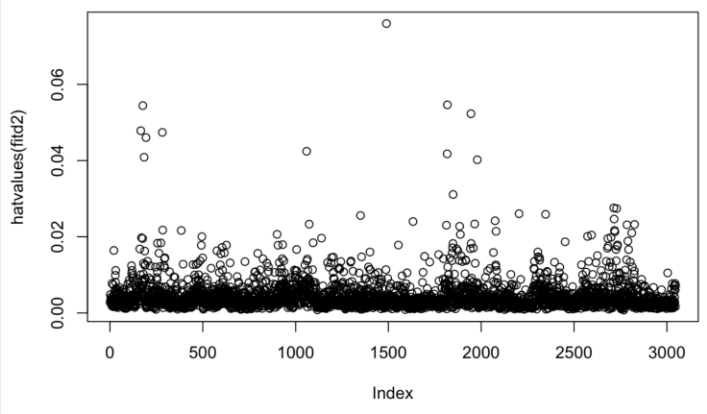
RE-FIT ON SELECTED VARIABLES, SOME OBSERVATIONS..


```
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              178.902845  15.240852  11.738  < 2e-16
## MedianAge                 -0.828682   0.105323  -7.868 4.97e-15
## AvgHouseholdSize         -13.092752   2.192771  -5.971 2.63e-09
## PctHS18_24                 0.307979   0.046215   6.664 3.15e-11
## PctHS25_Over               0.409123   0.092694   4.414 1.05e-05
## PctBachDeg25_Over         -1.112476   0.137121  -8.113 7.09e-16
## PctUnemployed16_Over       0.530894   0.156159   3.400 0.000683
## PctPrivateCoverage        -0.403922   0.094881  -4.257 2.13e-05
## PctWhite                  -0.102311   0.029314  -3.490 0.000489
## PctOtherRace              -0.879265   0.118489  -7.421 1.51e-13
## povertyPercent             0.149029   0.126172   1.181 0.237632
## PctPublicCoverageAlone     0.209718   0.143807   1.458 0.144853
## incidenceRate              0.195023   0.007126  27.367  < 2e-16
## BirthRate                 -0.952545   0.183936  -5.179 2.38e-07
```


```
## Residual standard error: 19.49 on 3033 degrees of freedom
## Multiple R-squared:  0.5088, Adjusted R-squared:  0.5067
## F-statistic: 241.7 on 13 and 3033 DF,  p-value: < 2.2e-16
```

DIAGNOSTIC PLOTS & CONCLUSIONS:





THANK YOU