STAT 6021 Project2 (Group 2) Proposal: Characteristics of nursing homes in New Mexico.

Gregory Madden, Christina Kuang, Chi Do, Trey Hamilton

Part 2: Proposal

Names of the R package and the R dataframe:

R Package: "data("Nursing")"

R dataframe: "Nursing"

Description of Dataset (provided by Stat2Data Package)

Characteristics of nursing homes in New Mexico.

Format: A dataset with 52 observations on the following 7 variables.

Beds Number of beds in the nursing home InPatientDays Annual medical in-patient days (in hundreds) AllPatientDays Annual total patient days (in hundreds) PatientRevenue Annual patient care revenue (in hundreds of dollars) NurseSalaries Annual nursing salaries (in hundreds of dollars) FacilitiesExpend Annual facilities expenditure (in hundreds of dollars) Rural 1=rural or 0=non-rural

Details: The data were collected by the Department of Health and Social Services of the State of New Mexico and cover 52 of the 60 licensed nursing facilities in New Mexico in 1988.

References: Howard L. Smith, Niell F. Piland, and Nancy Fisher, "A Comparison of Financial Performance, Organizational Characteristics, and Management Strategy Among Rural and Urban Nursing Facilities," Journal of Rural Health, Winter 1992, pp 27-40.

A link to the data set:

Data can be found at: http://lib.stat.cmu.edu/DASL/Datafiles/Nursingdat.html

Saving a file containing the data, as a .csv file.

```
# For some reason, Nursing dataset does not call correctly when knitting
# markdown so commented here uncomment and run below code when running in new
# environment

# Data <- Nursing write.csv(Data, 'Nursing.csv')

Data <- read.csv("Nursing.csv")
Data <- subset(Data, select = -c(X))  #remove column X
```

Showing the first few entries of the dataset:

```
head(Data, 1)
```

```
##   Beds InPatientDays AllPatientDays PatientRevenue NurseSalaries
## 1  244           128            385          23521          5230
##   FacilitiesExpend Rural
## 1             5334     0
```

• Project objectives/goals. What questions is the group trying to answer, as well as potential practical implications (the more interesting and/or practical, the better) of the results?

Q1. What characteristics of nursing homes in New Mexico dictate annual nurse salaries at those institutions?

Practical implications of a linear model for predicting nurse salaries could be used by policymakers to rationally distribute subsidy funds to institutions that are expected to have the lowest salaries.

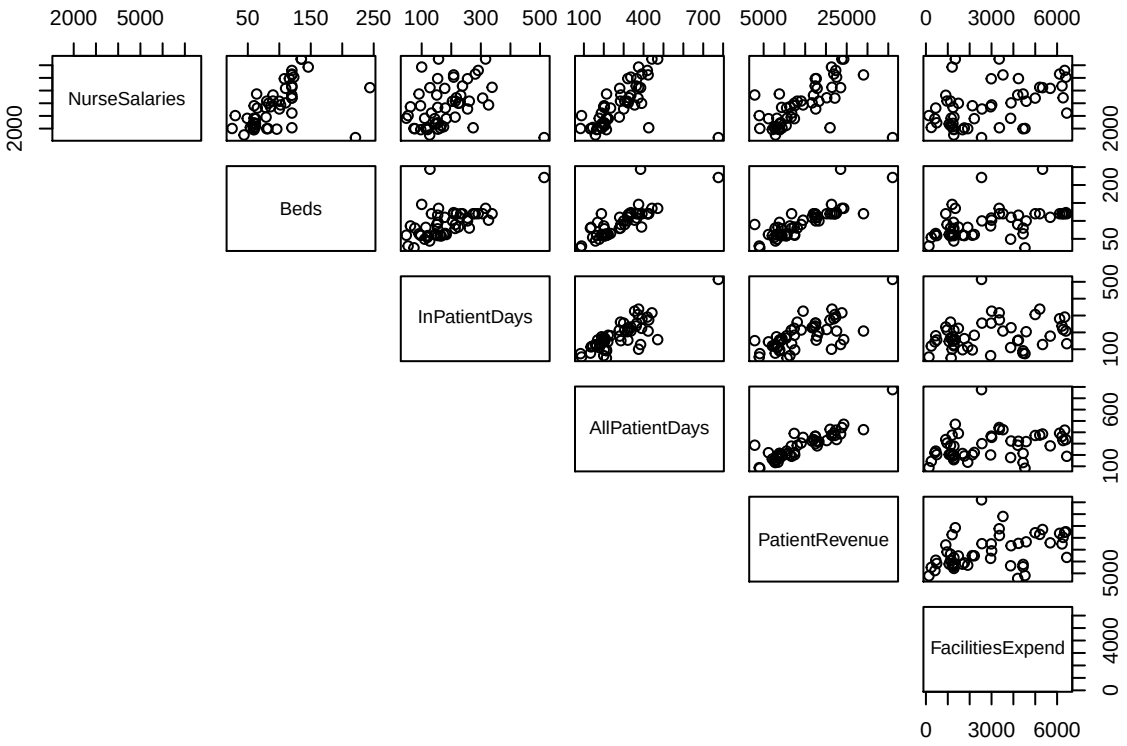
Q2. What factors are associated with rural vs. non-rural nursing homes?

Rural patients suffer from a lack of locally available nursing home beds. Understanding the relationships of these characteristics and how they define rural vs. non-rural nursing homes is helpful to know how to make rural nursing homes financially viable.

• Some data visualizations and commentary related to the project objectives/goals.

Based on the scatter plots and correlation table, it appears that Nurse Salaries has a moderate correlation with Beds, All Patient Days, and Patient Revenue. There also appears to be a strong linear relationship between Beds and AllPatientDays, Beds and Patient Revenue, In Patient Days and All Patient Days, In Patient Days and Patient Revenue, and All Patient Days and Patient Revenue. Further analysis is needed to determine if there is multicollinearity among the predictors.

```
pairs(~NurseSalaries + Beds + InPatientDays + AllPatientDays + PatientRevenue + FacilitiesExpend,
    data = Data, lower.panel = NULL)
```



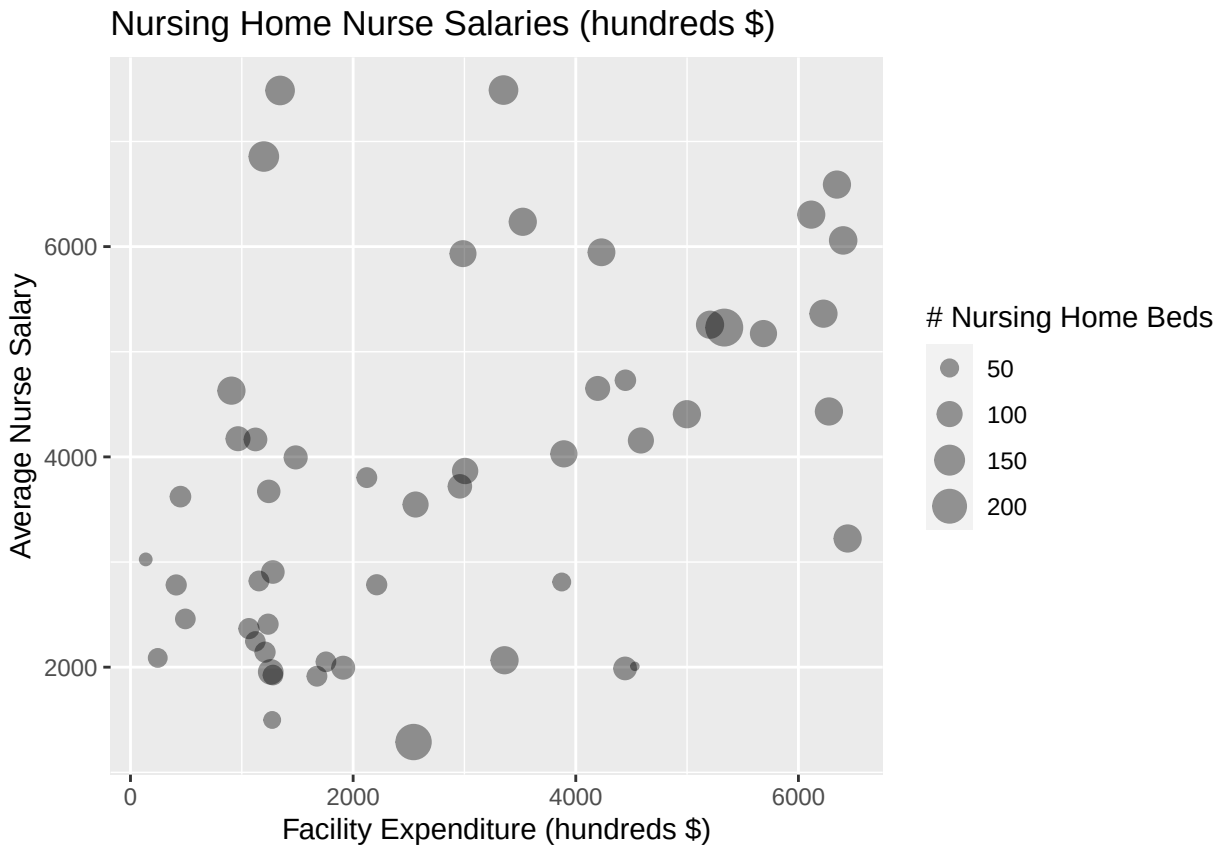

```
cor(Data[1:6])
```

```
##                     Beds InPatientDays AllPatientDays PatientRevenue
## Beds           1.0000000     0.5680006      0.8182959      0.8437752
## InPatientDays  0.5680006     1.0000000      0.8116225      0.7070754
```

```
## AllPatientDays    0.8182959     0.8116225      1.0000000      0.9030608
## PatientRevenue    0.8437752     0.7070754      0.9030608      1.0000000
## NurseSalaries     0.5094241     0.2541355      0.5153965      0.5894065
## FacilitiesExpend 0.4602559     0.2583959      0.3047354      0.4337859
##                   NurseSalaries FacilitiesExpend
## Beds                  0.5094241        0.4602559
## InPatientDays         0.2541355        0.2583959
## AllPatientDays        0.5153965        0.3047354
## PatientRevenue        0.5894065        0.4337859
## NurseSalaries         1.0000000        0.4550656
## FacilitiesExpend      0.4550656        1.0000000
```

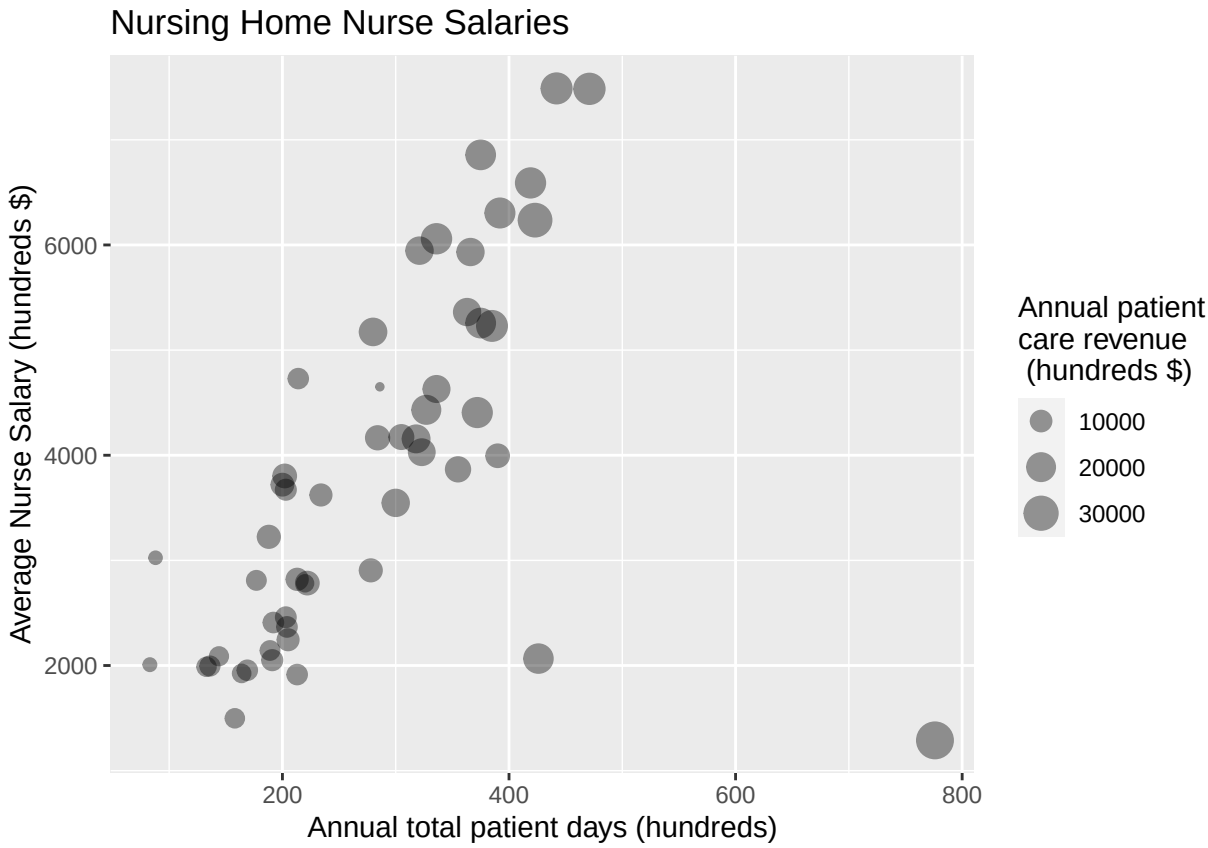
Nursing home salaries by Annual facilities expenditure (in hundreds of dollars) and Nursing home size (by number of beds): There appears to be a weak correlation between annual facility expenditure and average nurse salary. It also appears that facilities with more beds tend to have a higher expenditure and a higher average nurse salary.

```
Data %>%
    ggplot(aes(x = FacilitiesExpend, y = NurseSalaries, size = Beds)) + geom_point(alpha = 0.4) +
    labs(x = "Facility Expenditure (hundreds $)", y = "Average Nurse Salary", title = "Nursing Home Nurs
    guides(size = guide_legend(title = "# Nursing Home Beds"))
```



Nursing home salaries by census in Annual total patient days (in hundreds) and Nursing home size (by number of beds): There appears to be a moderately strong correlation between annual total patient days and average nurse salary. In addition, it also appears that facilities with higher total patient days tend to have a higher annual patient care revenue and a higher average nurse salary.

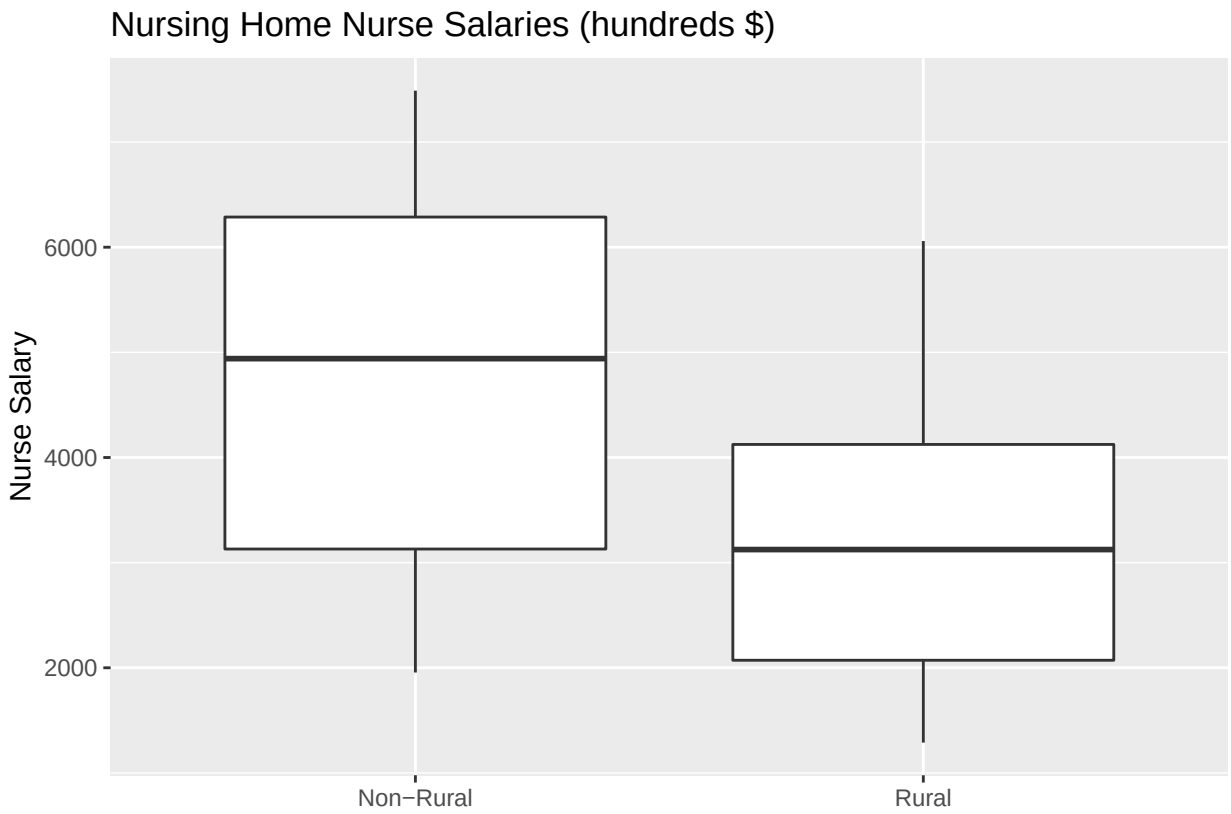
```
Data %>%
    ggplot(aes(x = AllPatientDays, y = NurseSalaries, size = PatientRevenue)) + geom_point(alpha = 0.4)
    labs(x = "Annual total patient days (hundreds)", y = "Average Nurse Salary (hundreds $)",
        title = "Nursing Home Nurse Salaries") + guides(size = guide_legend(title = "Annual patient \nca
```



Boxplot demonstrating differences in Institutional nurse's salaries in New Mexico for Rural Areas compared with Non-Rural: Based on the box plot, it appears there is a greater variability for non-rural nurse salary. The nurses in non-rural regions also have a higher median salary.

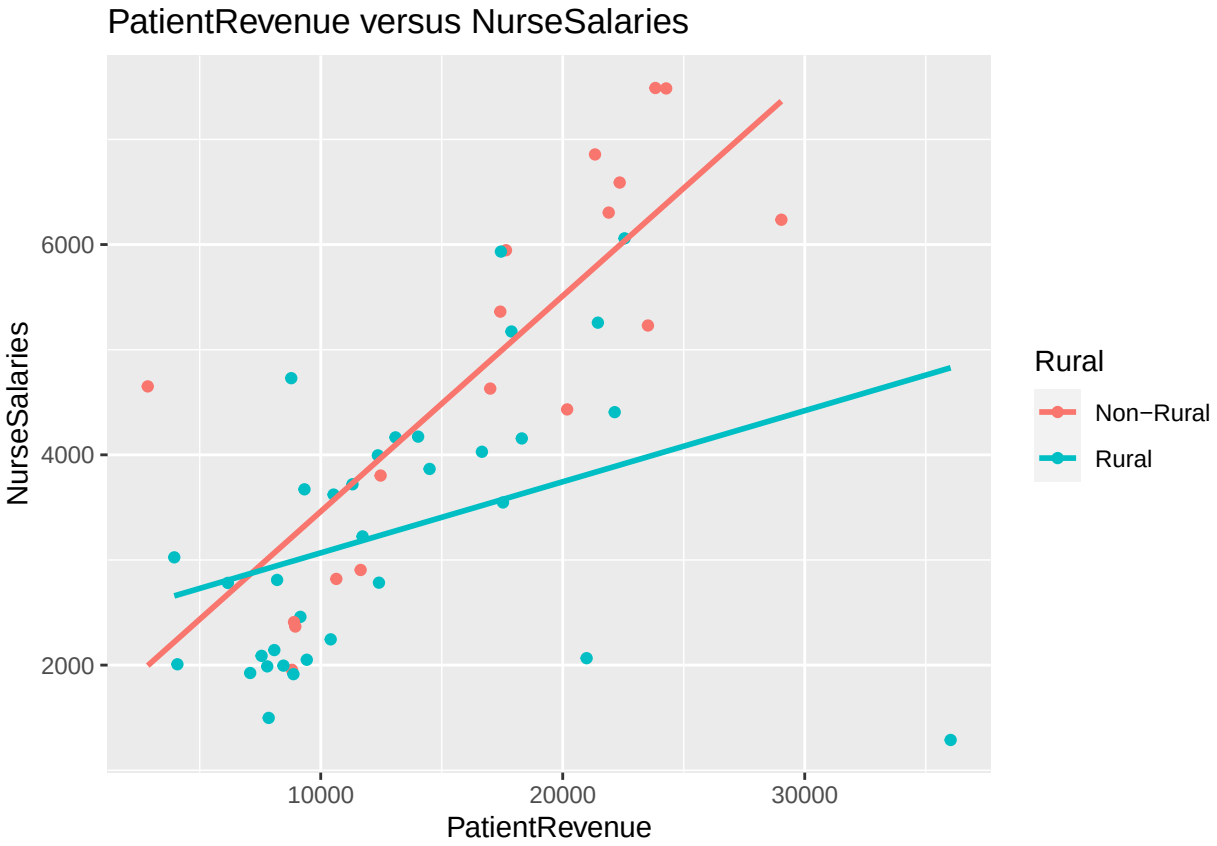
```
Data$Rural <- factor(Data$Rural)
levels(Data$Rural) <- c("Non-Rural", "Rural")

Data %>%
    ggplot(aes(x = Rural, y = NurseSalaries)) + geom_boxplot() + labs(x = "", y = "Nurse Salary",
    title = "Nursing Home Nurse Salaries (hundreds $)")
```



Scatter plot of Patient Revenue versus Nurse Salaries: The slopes are not parallel, which indicates there is an interaction effect between Patient Revenue and Nurse Salaries.

```
ggplot(Data, aes(x = PatientRevenue, y = NurseSalaries, color = Rural)) + geom_point() +
    geom_smooth(method = "lm", se = FALSE) + labs(x = "PatientRevenue", y = "NurseSalaries",
    title = "PatientRevenue versus NurseSalaries")
```



Objectives:

Objective 1: *Fit a multiple linear regression model with annual nurse salary for individual nursing homes using the available financial characteristics for each institution. The goal is to develop a model using these available data to reliably predict annual nursing salaries among the larger group of nursing homes across the state.*

Objective 2: *Fit a logistic regression model with rural vs. non-rural status as the response using the available financial characteristics.*