Predicting Rent Prices Using Socioeconomic and Demographic Predictors

David Lovejoy

1/9/2020

Executive Summary

This document outlines an analysis of monthly gross rent at the county level using various socioeconomic and demographic information. The census data was provided by the United States Department of Agriculture (USDA) Economic Research Service, and based on 40 specific features across all 50 states in the U.S.

After calculating summary and descriptive statistics, and creating visualizations to better understand the data, several potential relationships between renter characteristics and monthly gross rent were identified. After a thorough exploration the data, a linear regression model was used to predict monthly gross rent from renter features.

From the analysis of this admittedly small sample of a much larger general population, there were several significant features found in the health and demographic variables:

- **Health** — Higher instances of physical inactivity, obesity, and heart disease had a negative correlation with monthly gross rent.
- **Demographic** — Higher gross monthly rent tends to correlate with larger percentages of the population with at least a bachelor's degree. Conversely, there was a negative correlation with the percentage of the population with a high school diploma as the highest level of education, as well as with a higher death rate per 1,000 residents.

Data Exploration

Installing missing packages, importing the data and getting a sense for how it is structured:

```
data %>% as_tibble()
```

```
## # A tibble: 3,138 x 40
##    population renter_occupied~ pct_renter_occu~ evictions rent_burden
##         <int>            <int>            <dbl>     <int>       <dbl>
##  1      52842             5403             26.8        NA        28.0
##  2     212287            53502             57.5      6032        33.1
##  3      81263            13368             40.0      1012        32.0
##  4     122870            19359             41.9        NA        30.7
##  5     146153            15766             30.7       644        30.9
##  6      12511             1581             28.9         5        26.8
##  7       1334              134             18.6        NA        43.2
##  8      71635             6210             23.4        15        17.9
##  9     823964           117295             38.8      1841        31.2
## 10      11549             1170             21.7        26        31.6
## # ... with 3,128 more rows, and 35 more variables: pct_white <dbl>,
## #   pct_af_am <dbl>, pct_hispanic <dbl>, pct_am_ind <dbl>,
## #   pct_asian <dbl>, pct_nh_pi <dbl>, pct_multiple <dbl>, pct_other <dbl>,
```

```
## #   poverty_rate <dbl>, pct_civilian_labor <dbl>, pct_unemployment <dbl>,
## #   pct_uninsured_adults <dbl>, pct_uninsured_children <dbl>,
## #   pct_adult_obesity <dbl>, pct_adult_smoking <dbl>, pct_diabetes <dbl>,
## #   pct_low_birthweight <dbl>, pct_excessive_drinking <dbl>,
## #   pct_physical_inactivity <dbl>,
## #   air_pollution_particulate_matter_value <dbl>,
## #   homicides_per_100k <dbl>, motor_vehicle_crash_deaths_per_100k <dbl>,
## #   heart_disease_mortality_per_100k <int>, pop_per_dentist <int>,
## #   pop_per_primary_care_physician <int>, pct_female <dbl>,
## #   pct_below_18_years_of_age <dbl>, pct_aged_65_years_and_older <dbl>,
## #   pct_adults_less_than_a_high_school_diploma <dbl>,
## #   pct_adults_with_high_school_diploma <dbl>,
## #   pct_adults_with_some_college <dbl>,
## #   pct_adults_bachelors_or_higher <dbl>, birth_rate_per_1k <dbl>,
## #   death_rate_per_1k <dbl>, gross_rent <int>
```

There are 3,138 rows, in this case counties, and 40 columns, or features of each county. Next, a search and replacement for missing values was conducted.

```r
# check for missing values
sum(is.na(data))
```

```
## [1] 4972
```

```r
# replace missing values
data[is.na(data)] <- 0
# check again for missing values
sum(is.na(data))
```

```
## [1] 0
```

Partition the data into a train set and a test set:

```r
set.seed(1)
test_index <- createDataPartition(y = data$gross_rent, times = 1, p = 0.1, list = FALSE)
train <- data[-test_index,]
test <- data[test_index,]
```

Summarizing data in the target variable of **gross_rent**:

```r
summary(train$gross_rent)
```

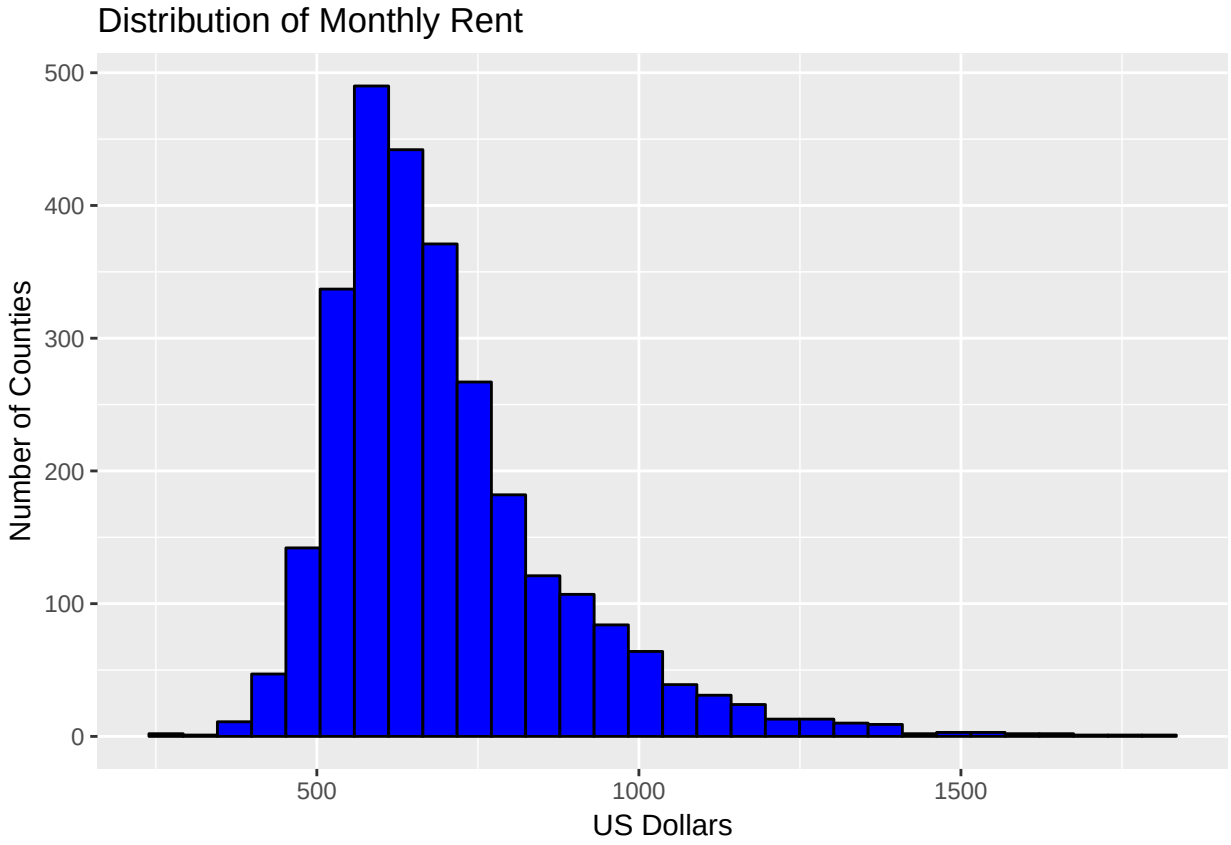
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   285.0   578.0   655.0   698.8   773.0  1827.0
```

```r
sd(train$gross_rent)
```

```
## [1] 182.5788
```

It is worth mentioning that the mean and median of the variable gross_rent, the focus of this analysis, differ somewhat significantly and that the relatively large standard deviation is indicative of considerable variance in the gross rent of different counties. A histogram of the gross_rent column shows that the rent values are right-skewed — meaning most rent values are located towards the lower end of the range, as shown here:

```
train %>%
ggplot(aes(gross_rent)) +
geom_histogram(bins = 30, fill = "blue", col = "black") +
xlab("US Dollars") +
ylab("Number of Counties") +
ggtitle("Distribution of Monthly Rent")
```



Checking variable correlation with **gross_rent**:

```
correlation <- apply(train,2, function(col)cor(col, train$gross_rent))
cor_table <- data.frame(correlation)
cor_table
```

```
##                                    correlation
## population                          0.41921023
## renter_occupied_households          0.35974744
## pct_renter_occupied                 0.30099061
## evictions                           0.25368531
## rent_burden                         0.23891498
## pct_white                          -0.25630574
## pct_af_am                           0.03652787
```

```
## pct_hispanic                                0.18757410
## pct_am_ind                                 -0.02247741
## pct_asian                                   0.59758349
## pct_nh_pi                                   0.19490857
## pct_multiple                                0.22930298
## pct_other                                   0.32371834
## poverty_rate                               -0.38249756
## pct_civilian_labor                          0.26913031
## pct_unemployment                           -0.11472396
## pct_uninsured_adults                       -0.18717075
## pct_uninsured_children                     -0.07598813
## pct_adult_obesity                          -0.47542235
## pct_adult_smoking                          -0.11026203
## pct_diabetes                               -0.45115890
## pct_low_birthweight                        -0.02011560
## pct_excessive_drinking                      0.29961926
## pct_physical_inactivity                    -0.58912691
## air_pollution_particulate_matter_value     -0.22240329
## homicides_per_100k                          0.12525487
## motor_vehicle_crash_deaths_per_100k        -0.31006095
## heart_disease_mortality_per_100k           -0.43937504
## pop_per_dentist                            -0.21960182
## pop_per_primary_care_physician             -0.17026728
## pct_female                                  0.01527041
## pct_below_18_years_of_age                   0.03933079
## pct_aged_65_years_and_older                -0.42694912
## pct_adults_less_than_a_high_school_diploma -0.32781247
## pct_adults_with_high_school_diploma        -0.60036817
## pct_adults_with_some_college                0.02051684
## pct_adults_bachelors_or_higher              0.69900272
## birth_rate_per_1k                           0.05363569
## death_rate_per_1k                          -0.62343239
## gross_rent                                  1.00000000
```

The percentage of adults with a bachelor's degree and the percentage of the Asian population both have a positive correlation with monthly rents. Conversely, the death rate per thousand residents, physical inactivity percentages, and percentages of adults with a high school diploma being the highest level of education received share a negative correlation with monthly rents. Interesting then, that there does not seem to be a strong correlation with the percentage of adults with less than a high school diploma.

Visualization

Create new data frame with potential predictors:

```
df <- data.frame(train$pct_asian,train$pct_adults_bachelors_or_higher,train$pct_adults_with_high_school

# rename columns
names(df)[names(df) == 'train.gross_rent'] <- 'Rent'
names(df)[names(df) == 'train.pct_physical_inactivity'] <- 'Inactivity'
names(df)[names(df) == 'train.pct_adults_with_high_school_diploma'] <- 'HS Diploma'
names(df)[names(df) == 'train.death_rate_per_1k'] <- 'Deaths'
names(df)[names(df) == 'train.pct_asian'] <- 'Asian Pop'
```

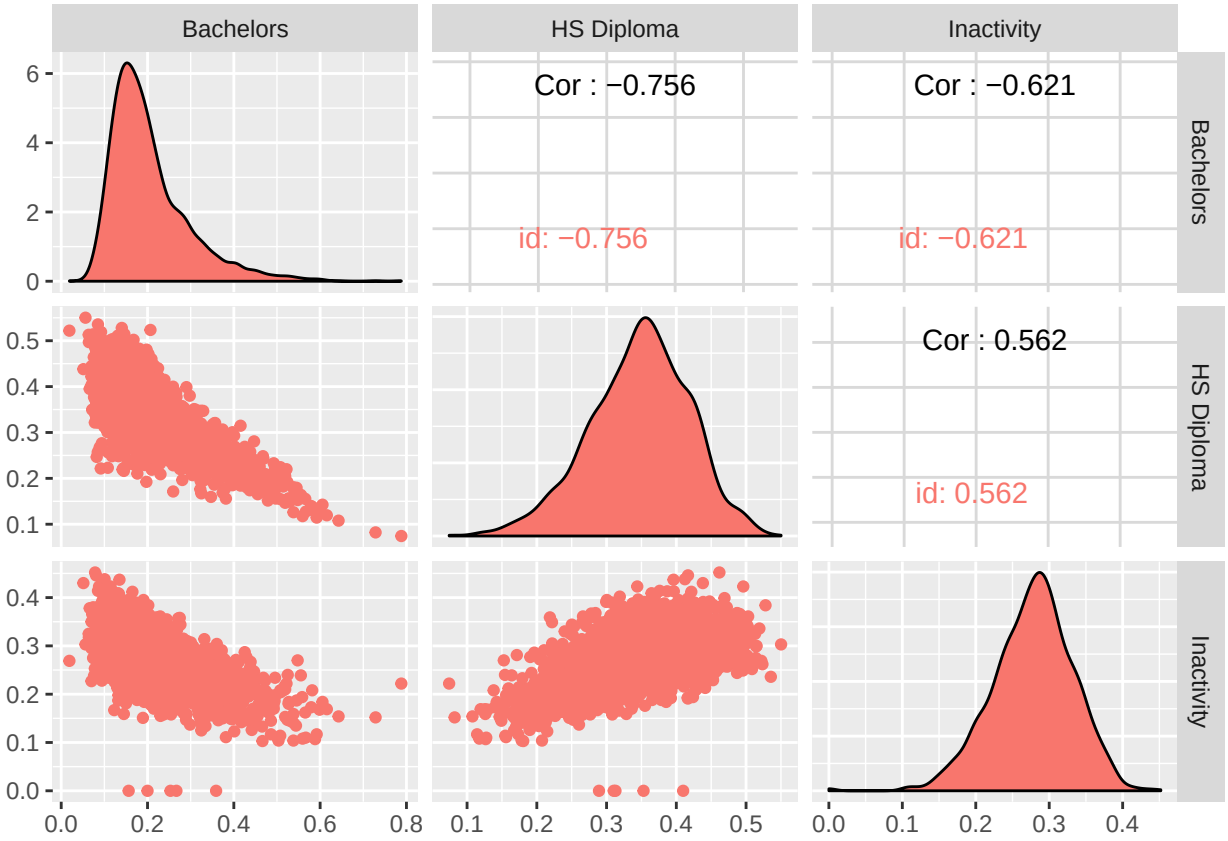
```
names(df)[names(df) == 'train.pct_adults_bachelors_or_higher'] <- 'Bachelors'

colnames(df)
```

```
## [1] "Asian Pop"  "Bachelors"  "HS Diploma" "Inactivity" "Deaths"
## [6] "Rent"
```

To better understand the complexity of the data, multi-faceted plots were made to discover and track patterns and relationships. There is perhaps a predictable inverse correlation between the percentage of the population obtaining a Bachelor's degree or higher and the percentage of the population with only a high school diploma.

```
ggpairs(df, columns = 2:4, ggplot2::aes(colour= "id"))
```

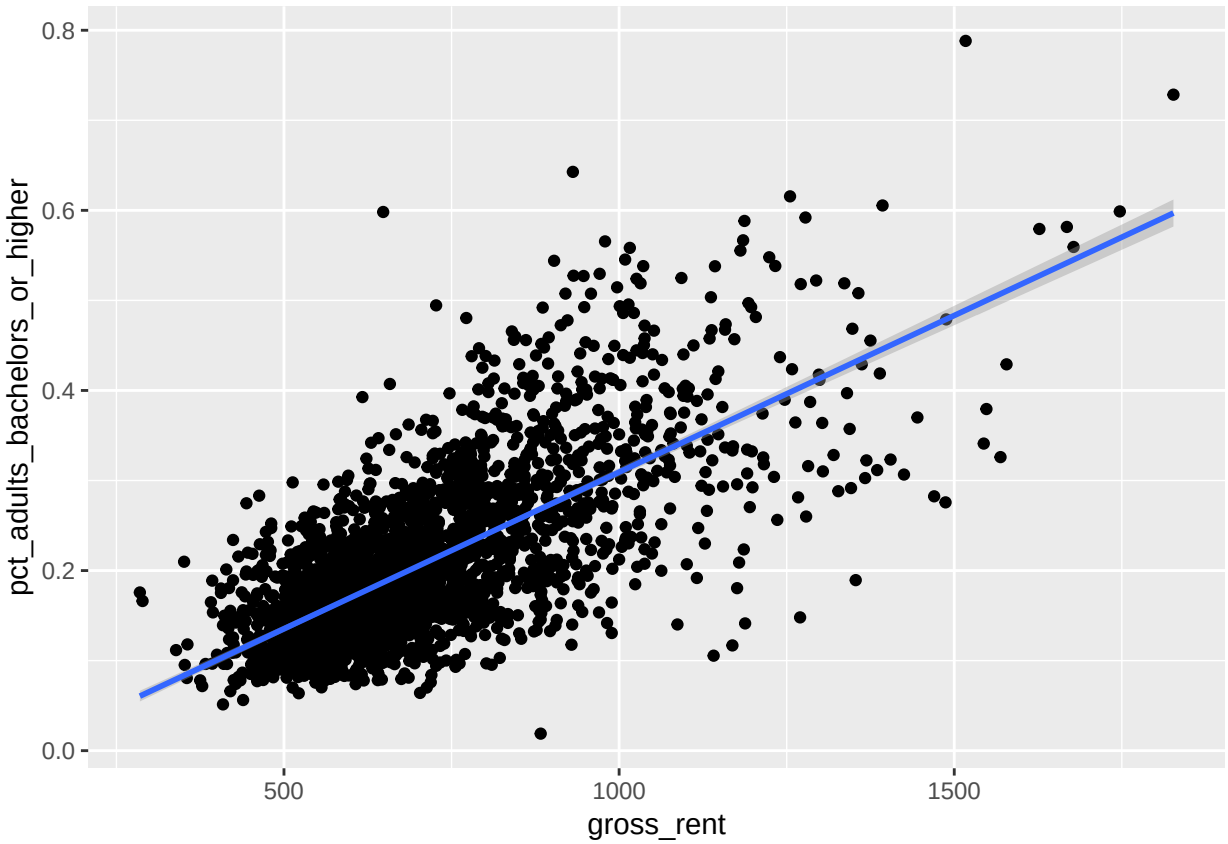


Rather surprisingly, physical inactivity and the percentages of the population with a high school diploma as the highest level of education received seem to be correlated.

Analysis

There seemed to be a clear linear relationship between the percentage of the population that has attained a bachelor's degree or higher level of education and the monthly rent prices.

```
train %>% ggplot(aes(gross_rent, pct_adults_bachelors_or_higher)) +
geom_point() +
geom_smooth(method = "lm")
```



Analysis

Based on the bivariate normal relationships identified when analyzing the data, a linear regression model was created to predict the value for monthly rent.

```
fit_1 <- lm(gross_rent ~ pct_adults_bachelors_or_higher, data = train)
```

Calculate the root mean squared error (RMSE) to see the accuracy of the prediction model on the test data.

```
prediction <- predict(fit_1, test)
error <- prediction - test[["gross_rent"]]
sqrt(mean(error^2))
```

```
## [1] 115.1329
```

That RMSE result does not look very promising. Try multiple regression model:

```
fit_2 <- lm(gross_rent ~ pct_adults_bachelors_or_higher + pct_asian + pct_adults_with_high_school_diplo

prediction <- predict(fit_2, test)
error <- prediction - test[["gross_rent"]]
sqrt(mean(error^2))
```

```
## [1] 101.4374
```

That is a sizeable improvement to the RMSE result, but still not ideal.

Conclusion

This analysis has shown that, while monthly rents at the county level are correlated with certain socioeconomic and demographic indicators, such as the percentages of the population physically inactivite and level of education completed, they cannot be predicted to a high level of accuracy using these variables alone. Furthermore, there is a surprising correlation between physical inactivity and percentages of the population to not have obtained a degree beyond a high school diploma, which should warrant further research.