Homework 1: Using OLS Regression to Predict Median House Values in Philadelphia

Yiming Cao, Sujan Kakumanu, Angel Sanaa Rutherford

2025-10-15

1 Introduction

2 Methods

2.1 Data Cleaning

The original dataset contained 1,816 Census block groups across Philadelphia. Following the data preparation protocol provided in the assignment, we removed four types of block groups:
1. Those with fewer than 40 residents,
2. Those with no housing units,
3. Those with median house values below $10,000, and
4. One extreme outlier in North Philadelphia with an exceptionally high median house value (over $800,000) and a very low median household income (below $8,000).

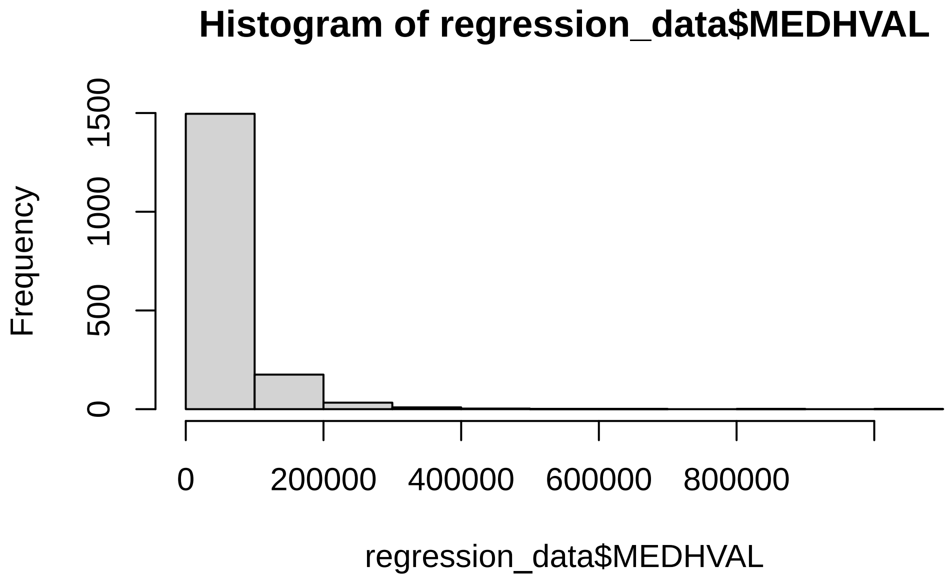
After applying these filters, the final cleaned dataset consisted of 1,720 block groups. All variables used in this analysis were numeric and contained no missing values after cleaning.

2.2 Exploratory Data Analysis

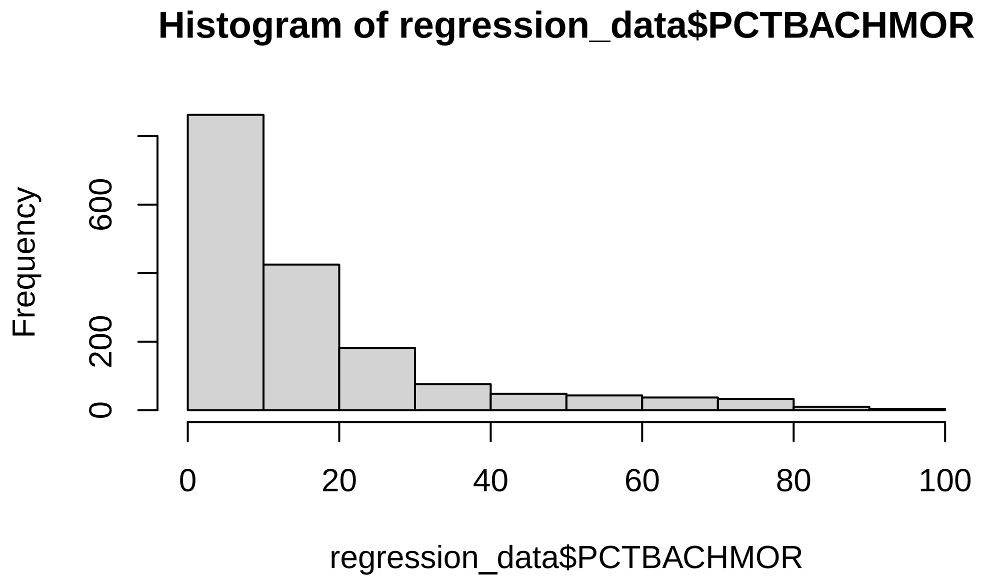
2.2.1 Variable Distributions

We began by importing the cleaned dataset and examining the distributions of the dependent variable (**MEDHVAL**) and four key predictors:
- **PCTBACHMOR** – Percentage of residents with at least a bachelor's degree
- **PCTVACANT** – Percentage of housing units that are vacant
- **PCTSINGLES** – Percentage of detached single-family homes
- **NBELPOV100** – Number of households below the poverty line

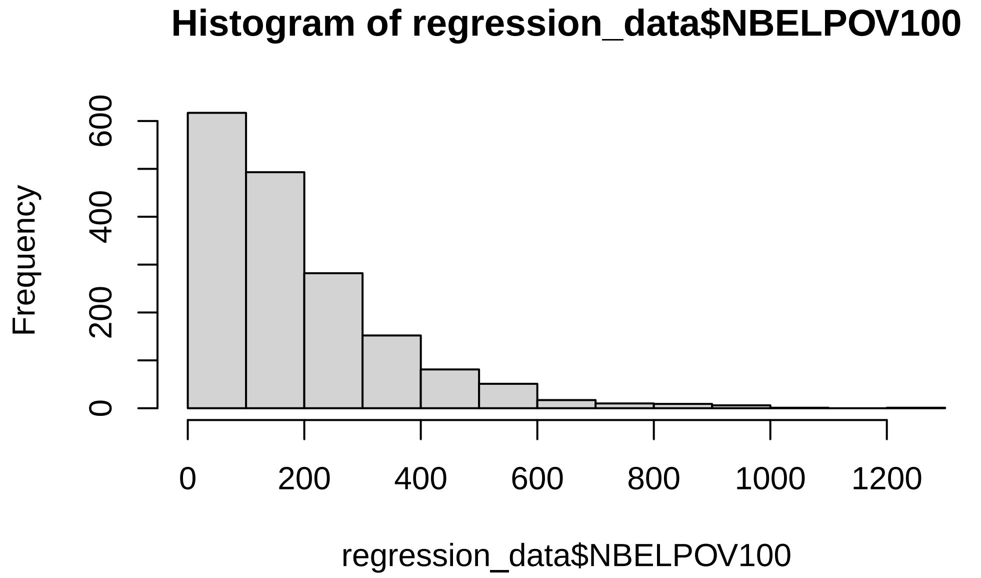
```
regression_data <- read.csv("./RegressionData.csv")

hist(regression_data$MEDHVAL)
```



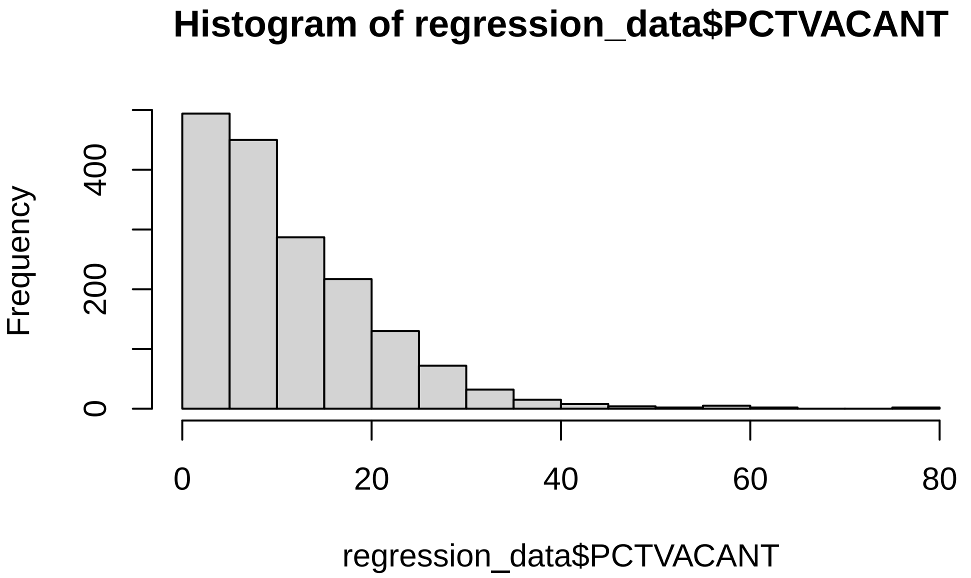
```
hist(regression_data$PCTBACHMOR)
```



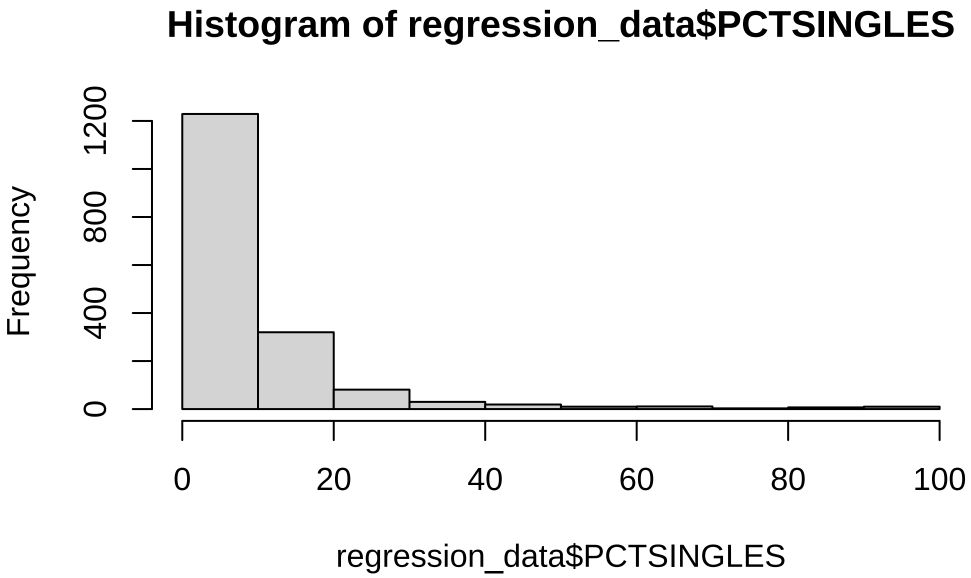
```r
hist(regression_data$NBELPOV100)
```



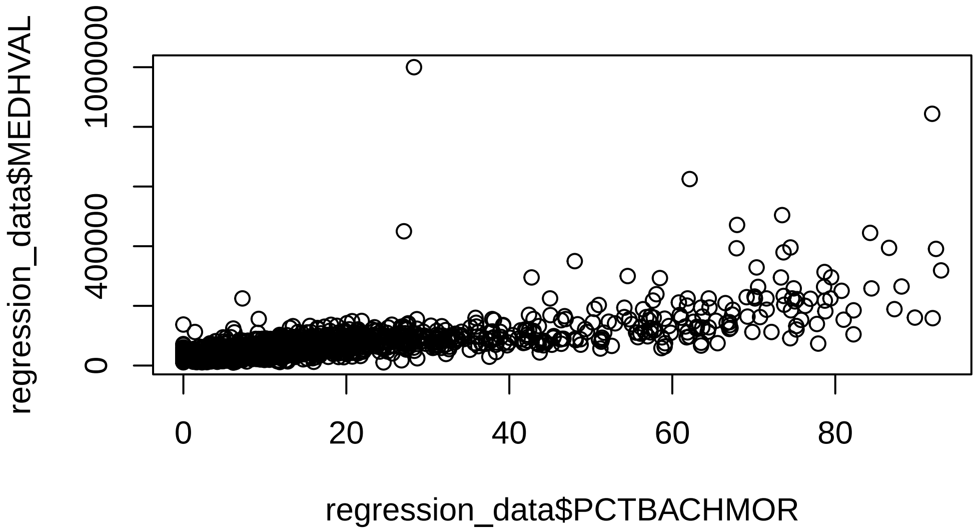
```r
hist(regression_data$PCTVACANT)
```



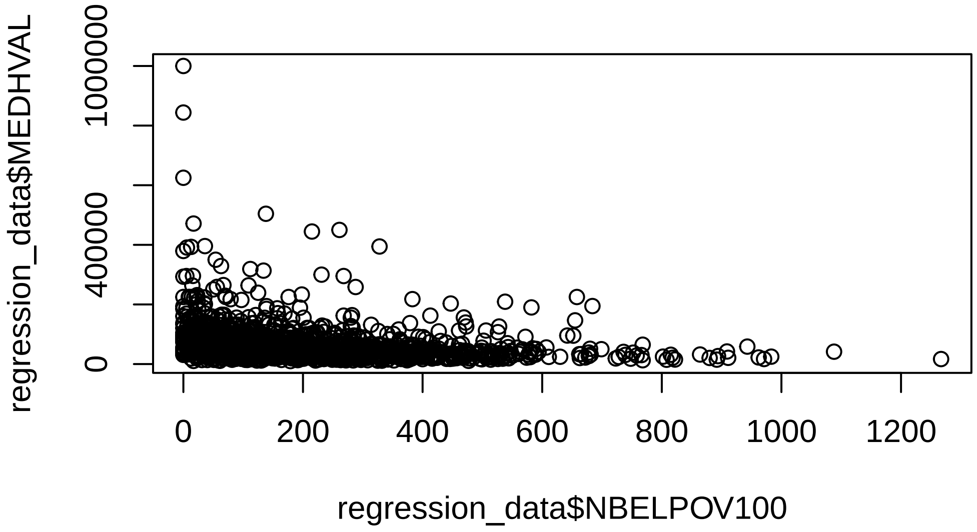
```r
hist(regression_data$PCTSINGLES)
```



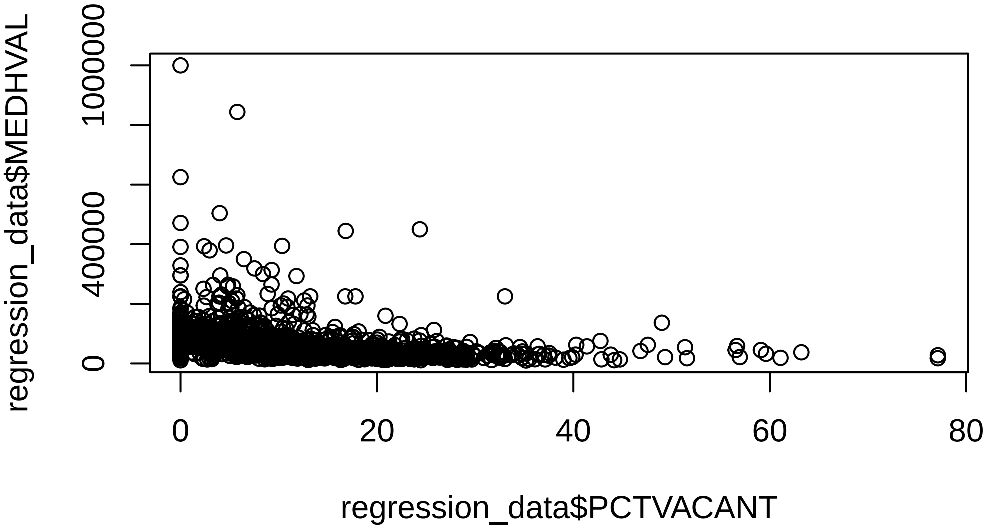
```
plot(regression_data$PCTBACHMOR, regression_data$MEDHVAL)
```



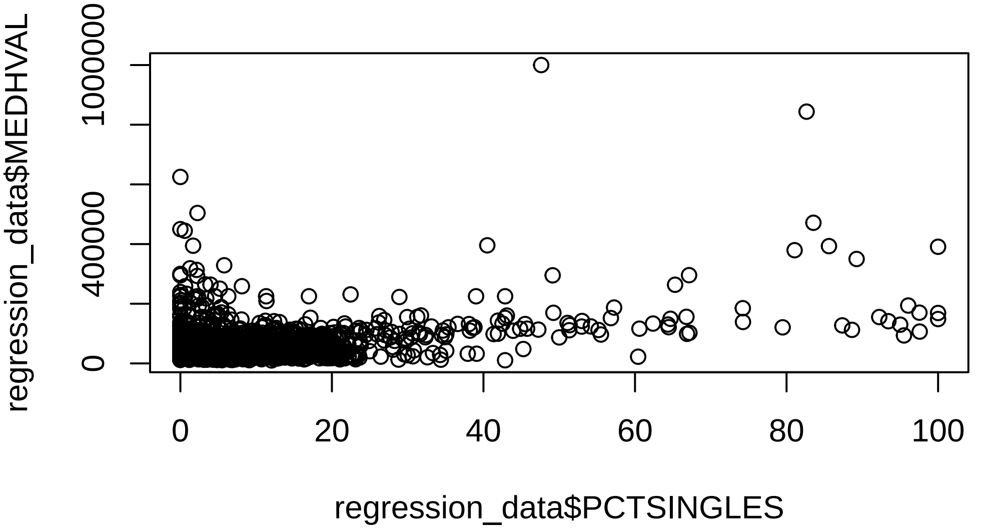
```
plot(regression_data$NBELPOV100, regression_data$MEDHVAL)
```



```r
plot(regression_data$PCTVACANT, regression_data$MEDHVAL)
```



```r
plot(regression_data$PCTSINGLES, regression_data$MEDHVAL)
```



```
mean_medhval <- mean(regression_data$MEDHVAL)
mean_pctbachmor <- mean(regression_data$PCTBACHMOR)
mean_nbelpov100 <- mean(regression_data$NBELPOV100)
mean_pctvacant <- mean(regression_data$PCTVACANT)
mean_pctsingles <- mean(regression_data$PCTSINGLES)

sd_medhval <- sd(regression_data$MEDHVAL)
sd_pctbachmor <- sd(regression_data$PCTBACHMOR)
sd_nbelpov100 <- sd(regression_data$NBELPOV100)
sd_pctvacant <- sd(regression_data$PCTVACANT)
sd_pctsingles <- sd(regression_data$PCTSINGLES)


# Create a tidy summary table
table1 <- tibble(
  Variable = c(
    "Median House Value (MEDHVAL)",
    "Households Below Poverty (NBELPOV100)",
    "% with Bachelor’s or Higher (PCTBACHMOR)",
    "% Detached Single-Family Homes (PCTSINGLES)",
    "% Vacant Housing Units (PCTVACANT)"
  ),
  Mean = c(66287.73, 189.77, 16.08, 9.23, 11.29),
  SD   = c(60006.08, 164.32, 17.70, 13.25, 9.63)
)

cat("Table 1. Summary statistics for dependent and predictor variables\n")
```

```
Table 1. Summary statistics for dependent and predictor variables
```

```
print(table1, row.names = FALSE)
```

```
# A tibble: 5 x 3
  Variable                                        Mean       SD
  <chr>                                          <dbl>    <dbl>
1 Median House Value (MEDHVAL)                66288.   60006.
2 Households Below Poverty (NBELPOV100)         190.     164.
3 % with Bachelor’s or Higher (PCTBACHMOR)       16.1     17.7
4 % Detached Single-Family Homes (PCTSINGLES)     9.23    13.2
5 % Vacant Housing Units (PCTVACANT)             11.3      9.63
```

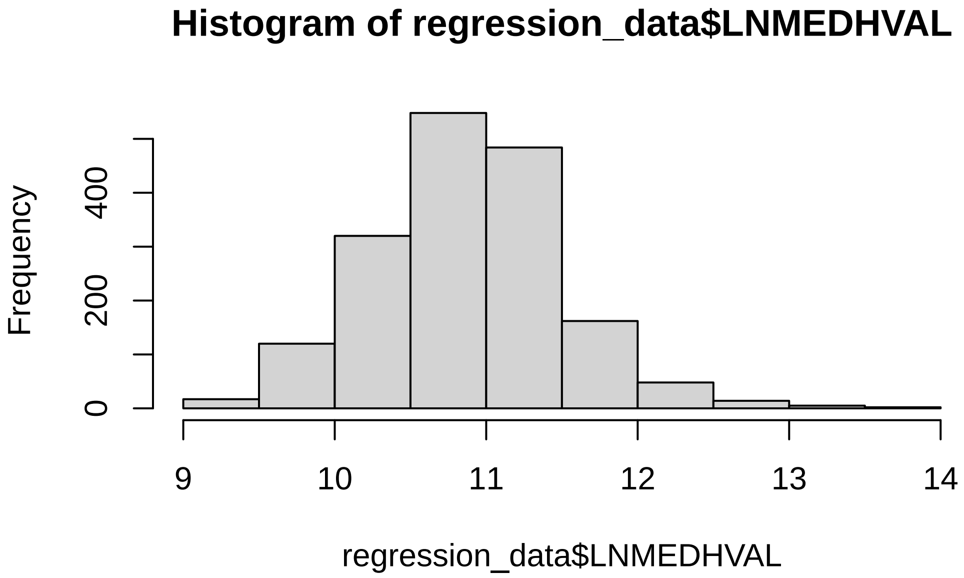
Summary statistics of the dependent and predictor variables are shown in Table 1.

All raw variables exhibited positive skewness, especially MEDHVAL and NBELPOV100, which justified a logarithmic transformation to stabilize variance.

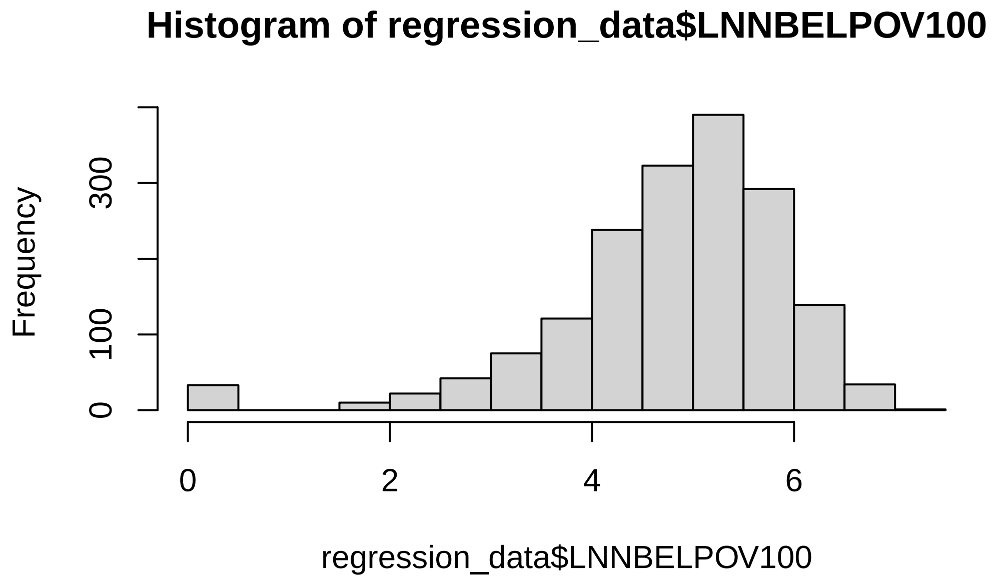
Log Transformations

The natural log of MEDHVAL was computed to improve normality, creating LNMEDHVAL as the dependent variable. For the predictors, only NBELPOV100 benefited from transformation, producing LNNBELPOV100. Other predictors retained strong spikes at zero after log transformation and were thus kept in their original form.

```r
regression_data$LNMEDHVAL<-log(regression_data$MEDHVAL)
regression_data$LNNBELPOV100<-log(1+regression_data$NBELPOV100)
regression_data$LNPCTBACHMOR<-log(1+regression_data$PCTBACHMOR)
regression_data$LNPCTVACANT<-log(1+regression_data$PCTVACANT)
regression_data$LNPCTSINGLES<-log(1+regression_data$PCTSINGLES)

hist(regression_data$LNMEDHVAL)
```



```r
hist(regression_data$LNNBELPOV100)
```



The log transformation substantially improved the symmetry of LNMEDHVAL, which appeared approximately normal. LNNBELPOV100 also showed improvement, whereas the log-transformed percentage variables remained zero-inflated.

Correlation Analysis

To examine potential multicollinearity, Pearson correlation coefficients were calculated among the four predictors. The sample correlation coefficient is defined as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

```r
trimmed_regression_data <- regression_data %>%
  dplyr::select(PCTBACHMOR, LNNBELPOV100, PCTVACANT, PCTSINGLES)

cor(trimmed_regression_data)
```

```
             PCTBACHMOR LNNBELPOV100  PCTVACANT PCTSINGLES
PCTBACHMOR    1.0000000   -0.3197668 -0.2983580  0.1975461
LNNBELPOV100 -0.3197668    1.0000000  0.2495470 -0.2905159
PCTVACANT    -0.2983580    0.2495470  1.0000000 -0.1513734
PCTSINGLES    0.1975461   -0.2905159 -0.1513734  1.0000000
```

Correlations were moderate: the strongest being between PCTBACHMOR and LNNBELPOV100 (r = –0.32). This indicates that the predictors were not highly collinear and all could be retained in the regression model.

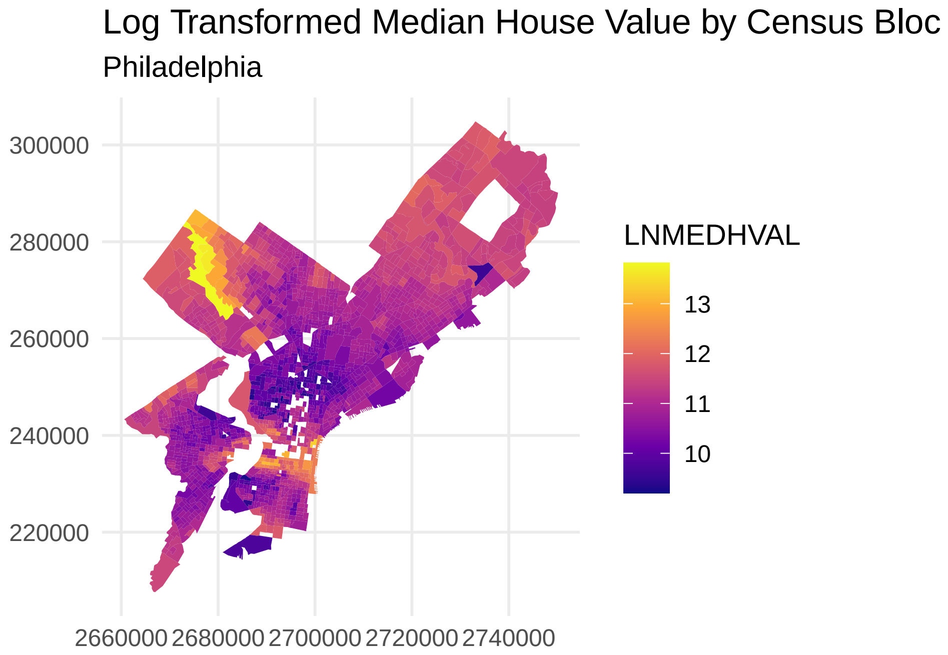
Mapping and Spatial Patterns

To visualize the spatial patterns, choropleth maps were created for LNMEDHVAL and each predictor variable using the shapefile provided.

```
Regression_shpData <- st_read("./Lecture 1 - RegressionData.shp")
```

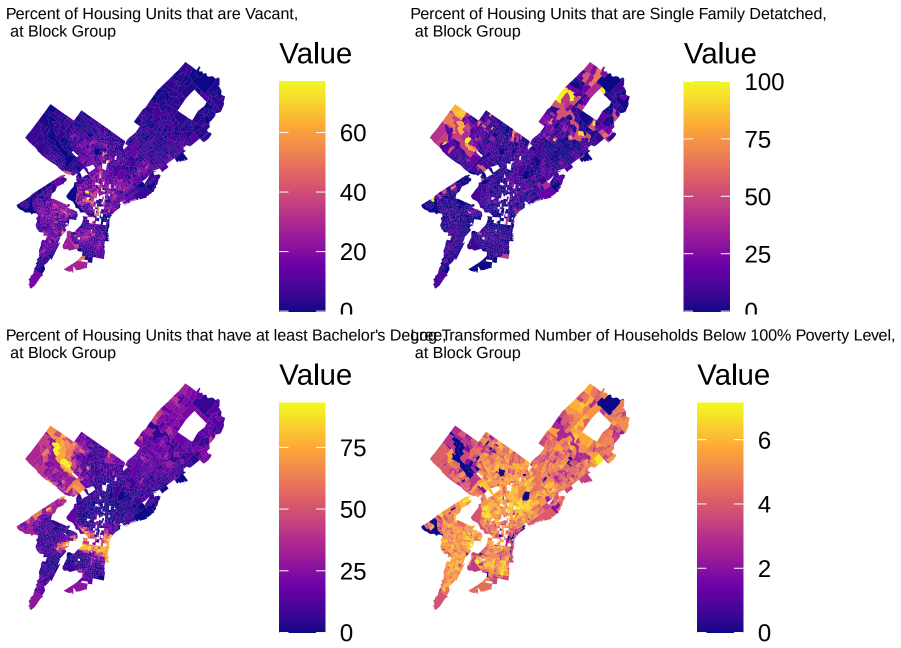
```
Reading layer `RegressionData' from data source
  `C:\Users\Yiming\Desktop\MUSA\5010 datamining\MUSA-5000\HW 1\Lecture 1 - RegressionData.shp
  using driver `ESRI Shapefile'
Simple feature collection with 1720 features and 13 fields
Geometry type: POLYGON
Dimension:     XY
Bounding box:  xmin: 2660605 ymin: 207610.6 xmax: 2750171 ymax: 304858.8
CRS:           NA
```

```
ggplot(Regression_shpData) +
  geom_sf(aes(fill = LNMEDHVAL), color = NA) +
  scale_fill_viridis_c(option = "plasma")+
  labs(title = "Log Transformed Median House Value by Census Block Groups", subtitle = "Phila
  theme_minimal()
```



```
p1 <- ggplot(Regression_shpData) +
  geom_sf(aes(fill = PCTVACANT), color = NA) +
  scale_fill_viridis_c(option = "plasma", na.value = "grey80") +
  labs(title = "Percent of Housing Units that are Vacant,\n at Block Group", fill = "Value")
  theme_minimal() +
  theme(axis.text = element_blank(), axis.title = element_blank(), panel.grid = element_blan
  base_theme

p2 <- ggplot(Regression_shpData) +
  geom_sf(aes(fill = PCTSINGLES), color = NA) +
  scale_fill_viridis_c(option = "plasma", na.value = "grey80") +
  labs(title = "Percent of Housing Units that are Single Family Detatched,\n at Block Group"
  theme_minimal() +
  theme(axis.text = element_blank(), axis.title = element_blank(), panel.grid = element_blan
  base_theme

p3 <- ggplot(Regression_shpData) +
  geom_sf(aes(fill = PCTBACHMOR), color = NA) +
  scale_fill_viridis_c(option = "plasma", na.value = "grey80") +
  labs(title = "Percent of Housing Units that have at least Bachelor's Degree,\n at Block Gro
  theme_minimal() +
  theme(axis.text = element_blank(), axis.title = element_blank(), panel.grid = element_blan
  base_theme
```

```
p4 <- ggplot(Regression_shpData) +
  geom_sf(aes(fill = LNNBELPOV), color = NA) +
  scale_fill_viridis_c(option = "plasma", na.value = "grey80") +
  labs(title = "Log Transformed Number of Households Below 100% Poverty Level,\n at Block Gro
  theme_minimal() +
  theme(axis.text = element_blank(), axis.title = element_blank(), panel.grid = element_blank
  base_theme

(p1 | p2) / (p3 | p4)
```


Percent of Housing Units that are Vacant,
at Block Group
Value
60
40
20
0
Percent of Housing Units that are Single Family Detatched,
at Block Group
Value
100
75
50
25
0
Percent of Housing Units that have at least Bachelor's Degree,
at Block Group
Value
75
50
25
0
Log Transformed Number of Households Below 100% Poverty Level,
at Block Group
Value
6
4
2
0

These maps revealed distinct spatial clustering: higher house values and educational attainment were concentrated in Center City and the northwest, while vacancy and poverty were highest in North and West Philadelphia. The pattern suggested potential spatial dependence, which will be tested in future assignments.

2.3 Multiple Regression Analysis

2.4 Additional Analysis

2.5 Software Used

3 Results

3.1 Exploratory Results

Table 1 summarizes the descriptive statistics for all key variables. LNMEDHVAL exhibited approximately normal distribution after transformation, while NBELPOV100 was best represented in logarithmic form. The percentage-based predictors remained right-skewed but were retained due to interpretability.

The correlation matrix indicated moderate relationships among predictors, with no evidence of severe multicollinearity. LNMEDHVAL correlated most strongly with PCTBACHMOR (r = 0.736) and negatively with PCTVACANT (r = –0.514) and LNNBELPOV100 (r = –0.424), confirming the visual impressions from the choropleth maps.

The spatial maps (Figure 1) clearly illustrated neighborhood-level variation in socioeconomic conditions.

High-value neighborhoods overlapped with areas of high educational attainment.

Poverty and vacancy rates were concentrated in North and West Philadelphia.

Detached single-family housing was more common in outer and suburban-edge block groups.

These exploratory results collectively indicate that higher educational attainment and single-family housing shares are positively associated with housing values, whereas higher vacancy and poverty rates are negatively associated. This provides a strong theoretical and empirical foundation for the multiple regression analysis that follows.

3.2 Regression Results

3.3 Regression Assumption Checks

3.4 Additional Models

4 Discussion & Limitations