Regression Modeling Project Report

Sathwik Garimella
Logan Kreutzberger
Sejun Song
Doehyung Kim

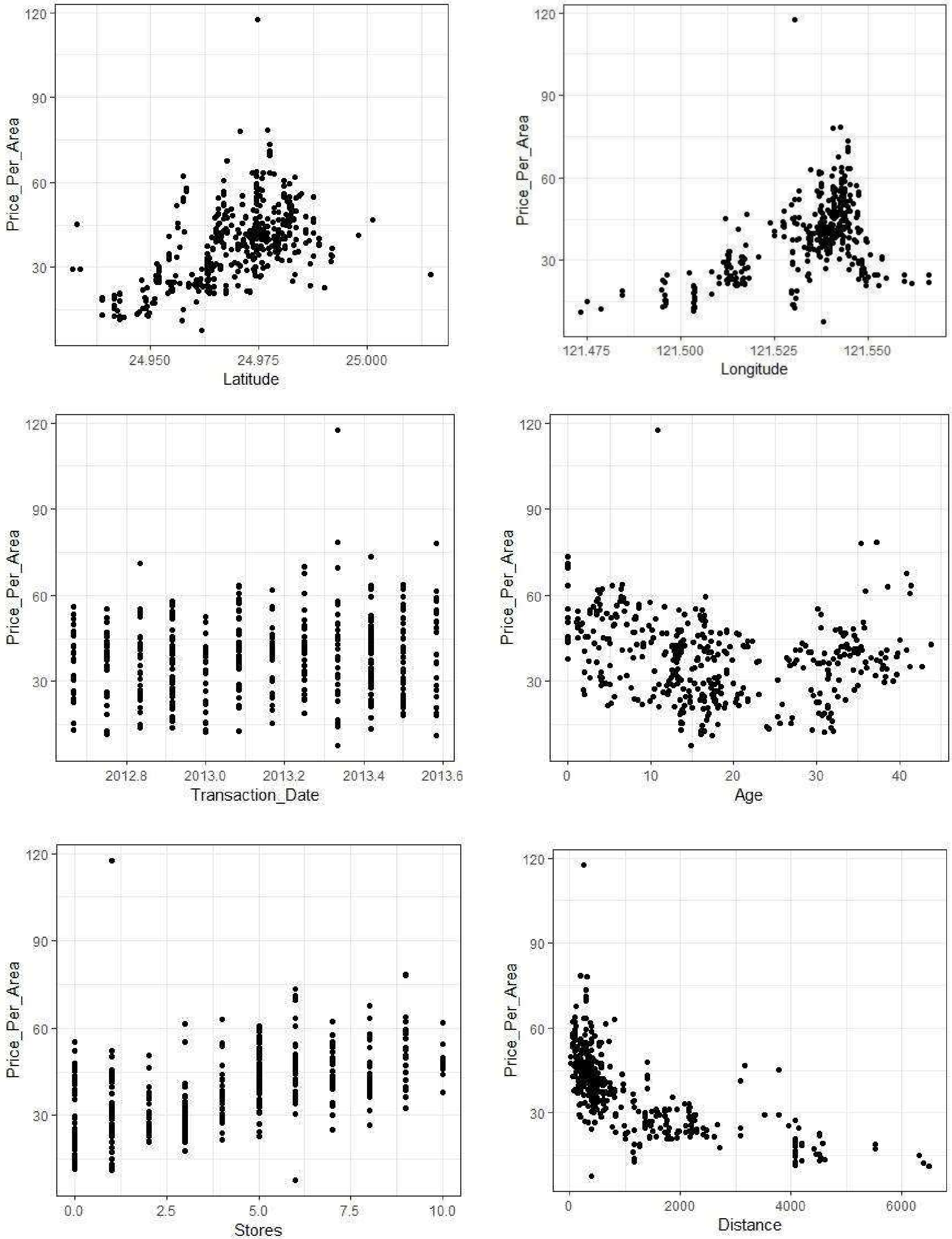
Introduction

- *Purchasing a house is a big decision in a person's life, and it takes a lot of consideration and research. People would like to buy a house at the best rate with minimum risk, and they wantit to be their best asset and investment for the future. In this project, we are going to analyze the data for House price prediction. We are going to predict the house prices to help people find the best house to invest using Linear Regression with factors such as transaction date of the house, house age, distance to the MTR station, number of convenience stores, latitude, and longitude of house.*

Methodology

- To start off the exploratory data analysis, we decided to see what our data looks like when put against the selected "y-variable." We decided to use the price per unit of area as our response variable as, out of the variablesin the dataset, it would likely be the most important factor for prospect homeowners. From here, we made predictions about whatvariables would be considered significant.
- After this, we fit the data into a model and summarized this model to see what variables were considered significant. This model was asimple additive model that contained all variables. Once this was finished, A new model was created with only significant variables. From here we also used the step method on the initial model to confirm that these variables were indeed significant. Usingthe summary function, we were also able to find values such as the p-value and R^2 value.
- *After this model was created, we checked thegraph of residuals versus the fitted line and the Q-Q plot. We did this to try to confirm normality. While normality and linearity were found to be satisfied, equal variance was not completely satisfied. To solve this problem, we took the log of the y-variable.*
- *This should have theoretically shrunk the variance to satisfy the conditions. This wasnot the case and variance still were found tonot be completely satisfied. Because of this, the reciprocal of the repone variable was taken. This solved the problem that there was with equal variance.*

Data: EDA


Price_Per_Area
Latitude
Longitude
Transaction_Date
Age
Stores
Distance

Model Analysis

Full Model

```
> model=lm(Price_Per_Area~Transaction_Date+Age+Distance+Stores+La
titude+Longitude, data=estate)
> summary(model)

Call:
lm(formula = Price_Per_Area ~ Transaction_Date + Age + Distance +

    Stores + Latitude + Longitude, data = estate)

Residuals:
    Min      1Q  Median      3Q     Max
-35.664  -5.410  -0.966   4.217  75.193

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.444e+04  6.776e+03  -2.131  0.03371 *
Transaction_Date  5.146e+00  1.557e+00   3.305  0.00103 **
Age              -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
Distance         -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
Stores            1.133e+00  1.882e-01   6.023 3.84e-09 ***
Latitude          2.255e+02  4.457e+01   5.059 6.38e-07 ***
Longitude        -1.242e+01  4.858e+01  -0.256  0.79829
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.858 on 407 degrees of freedom
Multiple R-squared:  0.5824,    Adjusted R-squared:  0.5762
F-statistic: 94.59 on 6 and 407 DF,  p-value: < 2.2e-16
```

Reduced Model

```
> model2=lm(Price_Per_Area~Transaction_Date+Age+Distance+Stores+L
atitude, data=estate2)
> summary(model2)

Call:
lm(formula = Price_Per_Area ~ Transaction_Date + Age + Distance +

    Stores + Latitude, data = estate)

Residuals:
    Min      1Q  Median      3Q     Max
-35.623  -5.371  -1.020   4.244  75.346

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.596e+04  3.233e+03  -4.936 1.17e-06 ***
Transaction_Date  5.135e+00  1.555e+00   3.303  0.00104 **
Age              -2.694e-01  3.847e-02  -7.003 1.04e-11 ***
Distance         -4.353e-03  4.899e-04  -8.887  < 2e-16 ***
Stores            1.136e+00  1.876e-01   6.056 3.17e-09 ***
Latitude          2.269e+02  4.417e+01   5.136 4.36e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.848 on 408 degrees of freedom
Multiple R-squared:  0.5823,    Adjusted R-squared:  0.5772
F-statistic: 113.8 on 5 and 408 DF,  p-value: < 2.2e-16
```

Parameter testing

```
> ##Parameter Testing
> mod_transaction=lm(Price_Per_Area~Age+Distance+Stores+Latitude+Longitud
e, data=estate)
> anova(mod_transaction,model)
Analysis of Variance Table

Model 1: Price_Per_Area ~ Age + Distance + Stores + Latitude + Longitude
Model 2: Price_Per_Area ~ Transaction_Date + Age + Distance + Stores +
    Latitude + Longitude
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    408 32790
2    407 31933  1    857.04 10.924 0.001034 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mod_distance<-lm(Price_Per_Area~Transaction_Date+Age+Stores+Latitude+Lon
gitude, data=estate)
> anova(mod_distance,model)
Analysis of Variance Table

Model 1: Price_Per_Area ~ Transaction_Date + Age + Stores + Latitude +
    Longitude
Model 2: Price_Per_Area ~ Transaction_Date + Age + Distance + Stores +
    Latitude + Longitude
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    408 34997
2    407 31933  1    3064.5 39.059 1.039e-09 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

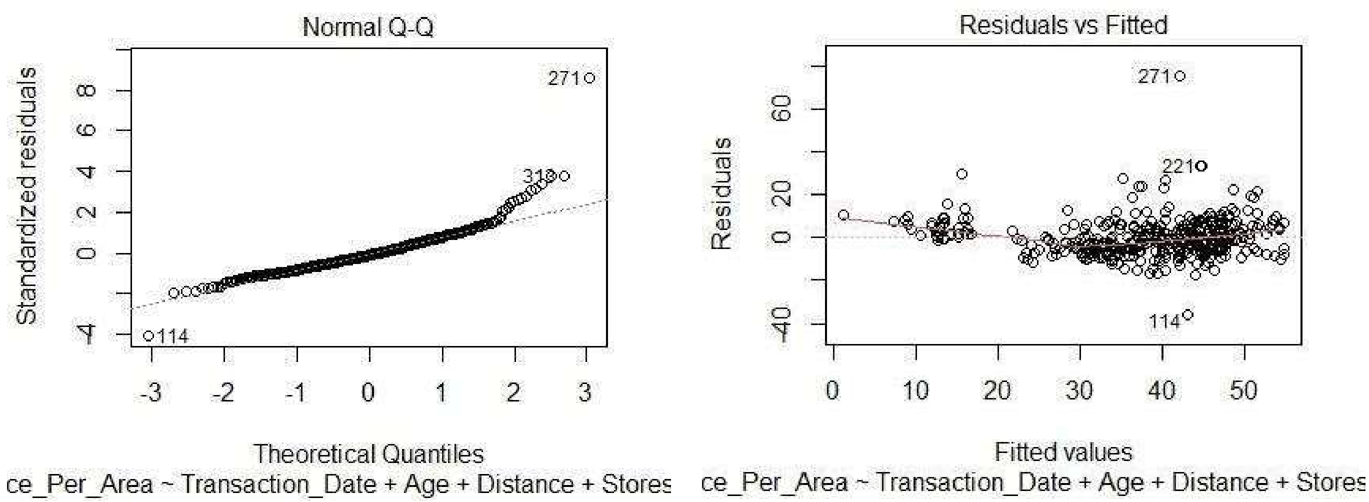
```
> mod_age<-lm(Price_Per_Area~Transaction_Date+Distance+Stores+Latitude+Lon
gitude, data=estate)
> anova(mod_age,model)
Analysis of Variance Table

Model 1: Price_Per_Area ~ Transaction_Date + Distance + Stores + Latitude
 +
    Longitude
Model 2: Price_Per_Area ~ Transaction_Date + Age + Distance + Stores +
    Latitude + Longitude
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    408 35776
2    407 31933  1    3843.9 48.993 1.065e-11 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mod_stores<-lm(Price_Per_Area~Transaction_Date+Age+Distance+Latitude+Lon
gitude, data=estate)
> anova(mod_stores,model)
Analysis of Variance Table

Model 1: Price_Per_Area ~ Transaction_Date + Age + Distance + Latitude +
    Longitude
Model 2: Price_Per_Area ~ Transaction_Date + Age + Distance + Stores +
    Latitude + Longitude
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    408 34779
2    407 31933  1      2846 36.274 3.835e-09 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> mod_latitude<-lm(Price_Per_Area~Transaction_Date+Age+Distance+Stores+Lon
gitude, data=estate)
> anova(mod_latitude,model)
Analysis of Variance Table

Model 1: Price_Per_Area ~ Transaction_Date + Age + Distance + Stores +
    Longitude
Model 2: Price_Per_Area ~ Transaction_Date + Age + Distance + Stores +
    Latitude + Longitude
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    408 33941
2    407 31933  1    2008.2 25.596 6.383e-07 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

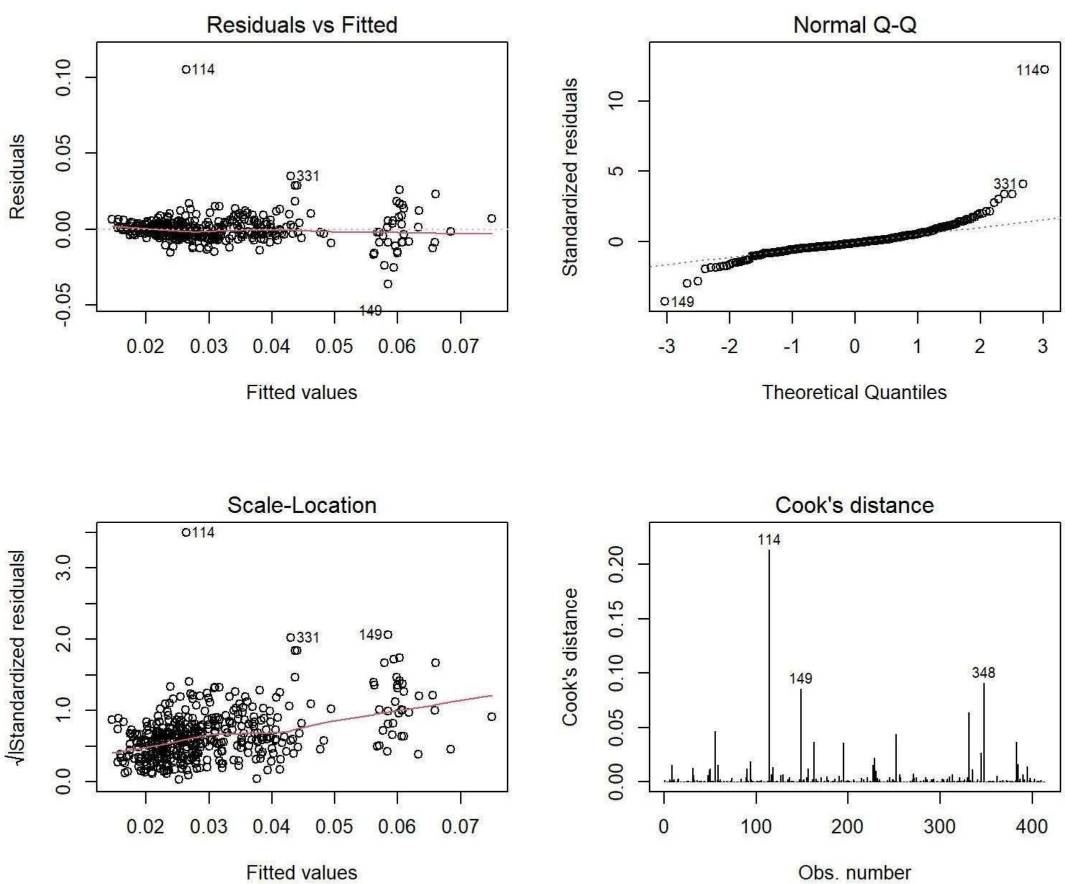
```
> mod_longitude<-lm(Price_Per_Area~Transaction_Date+Age+Distance+Stores+La
titude, data=estate)
> anova(mod_longitude,model)
Analysis of Variance Table

Model 1: Price_Per_Area ~ Transaction_Date + Age + Distance + Stores +
    Latitude
Model 2: Price_Per_Area ~ Transaction_Date + Age + Distance + Stores +
    Latitude + Longitude
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    408 31938
2    407 31933  1    5.1308 0.0654 0.7983
```

Residuals

Initial Residuals



Final Residuals

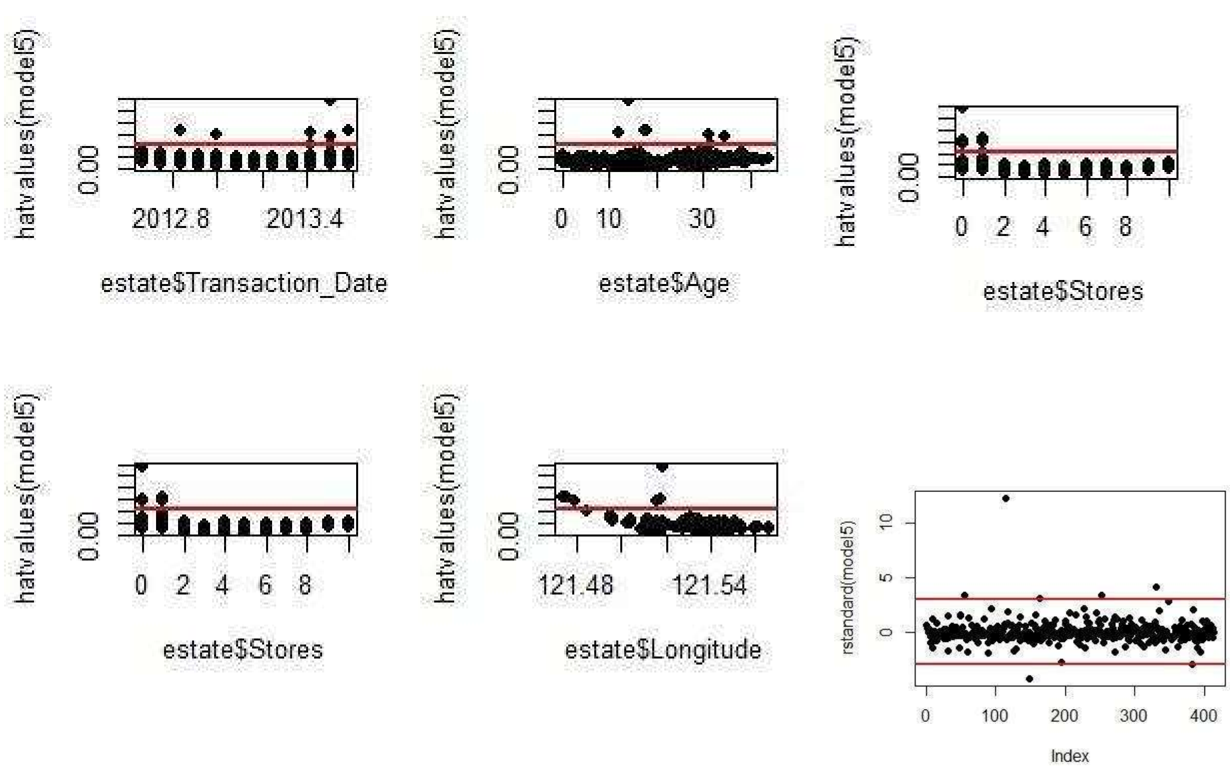


Normality Testing

```
> plot(model2, which=1:4)
> shapiro.test(residuals(model2))

        Shapiro-Wilk normality test

data:  residuals(model2)
W = 0.8755, p-value < 2.2e-16
```

```
> nortest::ad.test(residuals(model2))

        Anderson-Darling normality test

data:  residuals(model2)
A = 6.3907, p-value = 1.045e-15
```

Leverages



Conclusion

When we first did the EDA, we believed that age would be the variable that is insignificant due to the slight "u" shape that was made. This was not the case and the variable that ended up being insignificant was longitude. Because we were able to conclude that five of the six tested variables were significant, we would argue that the relationships found between these variables and the price per square foot of the area of a house would be useful to prospective homeowners when looking at houses to buy.

Based on the initial graphs, we were a bit surprised that age had a relationship when longitude did not as longitude appears to be more linear. The most challenging aspect of this project was the residuals. When we were able to fix equal variance, we lost some normality. We took leverages and studentized residuals to see if we could cut out any points. We decided not to cut out any points because many of the outliers for one variable would not be outliers for another point to an extent where it would be impossible to decide what points to cut out without disrupting the entire dataset. I think if we could redo this project, we would find a better dataset that has more non-significant variables. We also would like to find out what the missing 0.4 to the R-squared relates to.

A big downfall in this dataset was that the cost was based on units of area. We would have liked to see the unit of area as its own variable while the cost would also stand alone. Another thing that we noticed during the analysis process was how similar all the data were to one another. For example, the age range only consisted of houses built between 2012 and 2013 but was set up in a way factors could not apply. Longitude and latitude were similar with only houses at a longitude of 121.0-122.0 and latitude of 24.0-25.0 being measured. This made our data extremely catered to this specific area and age range which would not be beneficial for everyone.