# Project Deliverable 2

Yi Liu & Adam Kirstein
APAN 5200
Dr. Vishal Lala
Yi Liu_Adam Kirstein.R

## Brief Description of the Data:

The dataset on which we are conducting our analyses focuses on housing information collected from Craigslist.com. The dataset was scraped by a research team at the University of Washington in Seattle on October first, and includes house listings from as far back as March of 2016. The team has been collecting data through Craigslist.com, among many other sources during an ongoing exploration into the behavior of rent across Seattle, in part to measure the impact of housing ordinances enacted by the Seattle Housing Authority.

The dataset itself is broken into columns that indicate various aspects of housing, such as: Listing title, Listing date, Listing month, Rent price, Square Feet, Latitude and Longitude, Match address, number of Bathrooms and Bedrooms. The data file consists of 48 columns, and 40,600 rows, (90.56MB).

## Questions:

With this dataset, we hope to explore the intricacies of rent behavior within the city of Seattle. To do so, we have developed questions that will lead our analytical efforts into this data. Below outline or intentions for analysis:

1) **What housing criteria can we use to best predict rent?**
2) **Which model will allow us to predict rent with the highest accuracy?**
3) **What recommendations could we make for customers with specific housing criteria requirement?**

This project has set out to explain the variability of rental prices within Seattle, Washington, and by what criteria are these rental prices most affected, and then to create a predictive model that best predicts rental pricing. Our hypothesis suggests that price variability is most correlated with the number of bedrooms and bathrooms, as well as the total square footage of the apartment.

### A) Identifying relationships, and variable selection

We utilized a correlation test to assess what variables shared the most in common with our independent variable, RentPrice. Through this assessment, we discovered that number of beds, bath, and squarefeet all share a significant relationship with the price of rent, meaning that as either of these above-stated variables increase, so too does rent price.Thus, we selected these variables to explore their predictability.

## B) Testing for predictability

### a) Single Linear Regression analysis

We wanted to see at what rate of predictability our selected variables had together when predicting the outcome of rent. We individually tested the predictability of our selected variables using a simple linear regression model. Doing this allowed us to see how each of our selected variables explained the variability within the data. We found that low individual rates for each variables.

We then wanted to see how each of these variables interacted together, as well as alongside our variable of interest, namely RentPrice.

### b) Multiple Linear Regression analysis

We moved to a multiple regression model. From this model, we found that together our variables explained more than they did apart, but were still low, with a final R-squared value of 54.85%. We knew our variables were significant given their reported statistical significance, so we are able to reject the null hypothesis that Beds, Baths, and Square footage do not affect rent price.

Because of this, we were effective at identifying a likely trend amongst the data, as well as sharing a high correlation value with our dependent variable. We tested the effect of adding more variables into the model, namely, latitude, longitude, and Listing date, but the difference in variance was negligible, as each had little to no impact at

explaining the dependent variable. Given that we had already selected the highest correlated variables identified through a correlation matrix, we wanted to maintain only their use, instead of overfitting the model with less-correlated variables, thus adding noise.

Regression model was able to show some relationship in predictability between our variables, but the model was not sufficient enough to reduce variability.

Additionally, we speculated that location plays a role in the accuracy of our model. But expressing location was difficult while utilizing the lat/long columns. However, we noticed that by utilizing a subsetting function within our regression model, *lm1 = lm(RentPrice~Sqft+Beds, data= subset(CL_new, matchAddress =="40th Ave NE"))*
we were able to isolate specific localities within the “matchAddress” column, that could be used to lower variability for prediction. They offered less variability, and more honed results in terms of their ability to predict rental price.

Despite this, to test our predictability,  we applied the predict() function to our primary model, for example: *data*(Sqft=1450, Beds=2, Baths=2), Which gave the result of rent price of **$3287.334**, which is an (unfortunate) highly accurate representation of what is to be expected based on the criteria.

In  this way, customers could better predict the rent price by their specific bedroom/bathroom, square feet needs when they search the apartment.

## Clustering

In addition to our predictive analysis, we opted to segment parts of our data in order to address the potential specific need of a customer seeking a place to live under a certain criteria. We opted to randomly select a housing combination, 3 bedrooms and 1 bathroom,  and subset this data for clustering. Thanks to the latitude and longitude data, we were able to make a graphical overview in Seattle, (Fig12 ). By presenting the plot, apartment seekers could have more preparation to find houses in the market.

## Conclusions

We conclude that predicting rent price is viable with this dataset, and that the most influential variables affecting the fluctuation in rent is Beds, Baths, and Squarefeet, and is highly effective when restricted to specific, hyper-locality. We think this is an appropriate solution into the further exploration into rent price preference matching.

We were able to demonstrate the significance between rent price and our selected variables. Though we identified a trend in the dataset, we were not able to reduce the variance to fit the model closer. We suspect this has to do with some of our variables being collinear, and that a potential solution to this would be to further adjust the data by inspecting it deeper for collinearity.

# Appendix

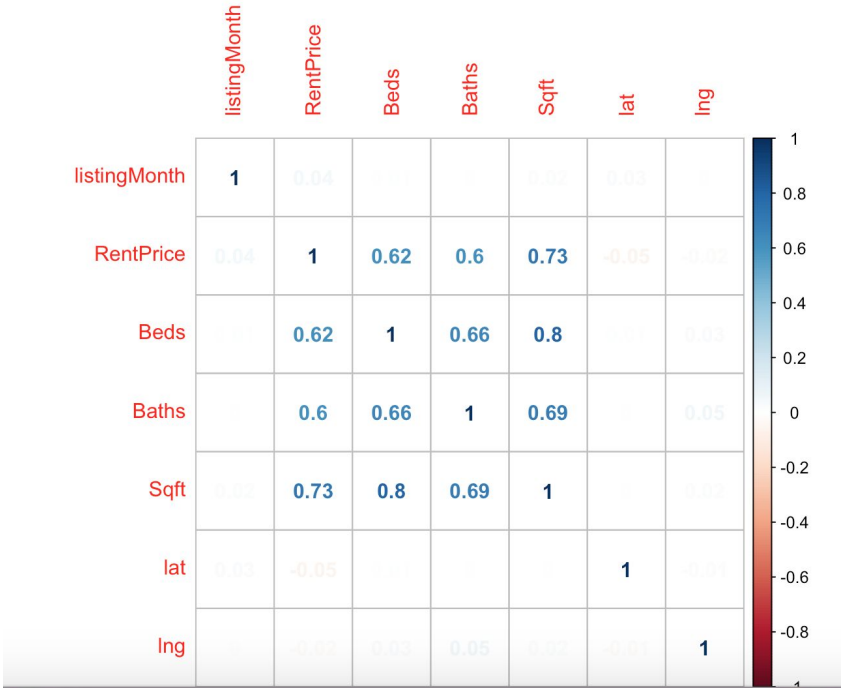
## Multiple Linear Regression

We use cor by corre.test to find the correlations for data and probability values.

```
> corr.test(CL_new[,c(3:7,10:11)])
Call:corr.test(x = CL_new[, c(3:7, 10:11)])
Correlation matrix
             listingMonth RentPrice Beds Baths Sqft   lat   lng
listingMonth         1.00      0.04 0.01  0.00 0.02  0.03  0.00
RentPrice            0.04      1.00 0.62  0.60 0.73 -0.05 -0.02
Beds                 0.01      0.62 1.00  0.66 0.80  0.01  0.03
Baths                0.00      0.60 0.66  1.00 0.69  0.00  0.05
Sqft                 0.02      0.73 0.80  0.69 1.00  0.00  0.02
lat                  0.03     -0.05 0.01  0.00 0.00  1.00 -0.01
lng                  0.00     -0.02 0.03  0.05 0.02 -0.01  1.00
Sample Size
[1] 35158
Probability values (Entries above the diagonal are adjusted for multiple tests.)
             listingMonth RentPrice Beds Baths Sqft  lat  lng
listingMonth         0.00         0 0.74  1.00 0.01 0.00 1.00
RentPrice            0.00         0 0.00  0.00 0.00 0.00 0.01
Beds                 0.12         0 0.00  0.00 0.00 1.00 0.00
Baths                0.67         0 0.00  0.00 0.00 1.00 0.00
Sqft                 0.00         0 0.00  0.00 0.00 1.00 0.01
lat                  0.00         0 0.32  0.95 0.87 0.00 0.31
lng                  0.72         0 0.00  0.00 0.00 0.04 0.00
```

**(Fig 1)**

From Correlation matrix, we can see that RentPrice has correlation with listingMonth(0.04), Beds(0.62), Baths(0.60), Sqft(0.73), lat(-0.05) and lng(-0.02). Therefore, we believed that RentPrice has strong relationship with Beds, Baths and Sqft. And in Probability values, their p-values are all less than 0.05. Therefore, we included variables of Beds, Baths, Sqft into our model.

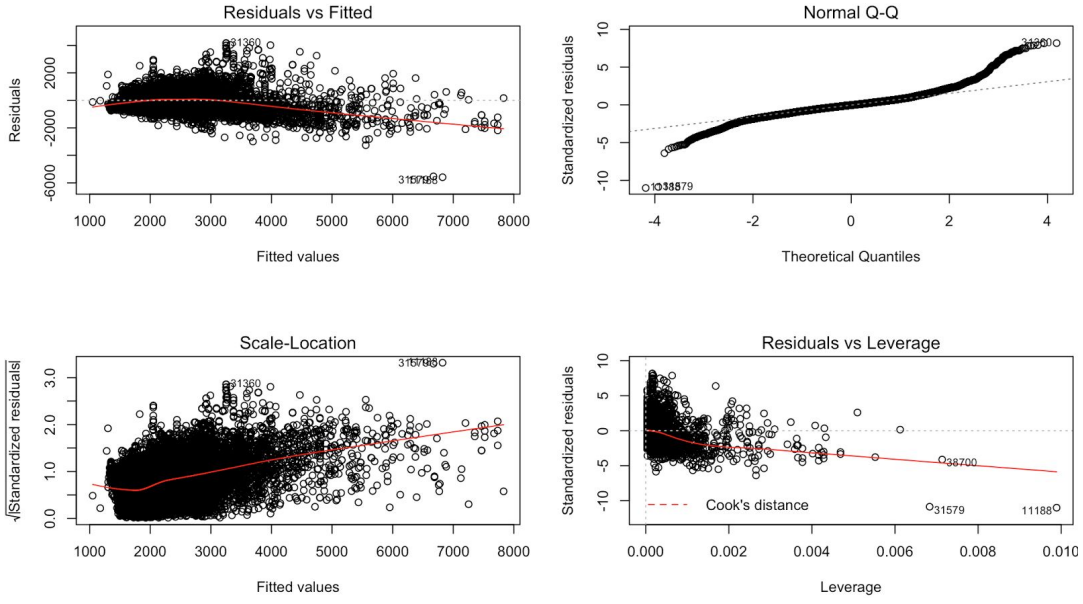



**(Fig 2)**

This corrplot Fig 2 displays the correlation coefficient vividly compared to the matrix. (**Fig 3)**

```
> model = lm(RentPrice~Sqft+Beds+Baths,data=CL_new)
> summary(model)

Call:
lm(formula = RentPrice ~ Sqft + Beds + Baths, data = CL_new)

Residuals:
    Min      1Q  Median      3Q     Max
-5592.3  -281.4   -19.8   251.5  4173.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 872.38761    8.30036 105.102  < 2e-16 ***
Sqft          1.20930    0.01357  89.132  < 2e-16 ***
Beds         32.11594    5.48937   5.851 4.94e-09 ***
Baths       298.61389    8.50443  35.113  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 510.7 on 35154 degrees of freedom
Multiple R-squared:  0.5485,    Adjusted R-squared:  0.5485
F-statistic: 1.424e+04 on 3 and 35154 DF,  p-value: < 2.2e-16
```

The minimum residuals is -5592.3, median -19.8, and maximum is 4173.9. The spread of Residuals looks like a normal distribution. R-squared is a goodness of fit. The adjusted R-squared=0.5485, which means under 54.85% situation, we could use this model to predict. Also, the p-value of the model is less than 0.05.
After fitting the regression model, we did Regression Diagnostics and plot the residuals by par(mfrow=c(2,2)), in a way to check if the model works well for data




.**(Fig 4)**

In Residuals vs Fitted plot, the spread residuals around a horizontal line without distinct patterns, so we don't have non-linear relationships. This Normal Q-Q shows the most residuals follow the straight dash line well, not bad residuals normality. In the Scale-Location plot, residuals randomly spread. However, in Residuals vs Leverage, we find some influential cases, such as 38700, 31579 and 11188 outside of the Cook's distance. If we exclude those cases, the regression results could be altered.

```
> model

Call:
lm(formula = RentPrice ~ Sqft + Beds + Baths, data = CL_new)

Coefficients:
(Intercept)         Sqft         Beds        Baths
    872.388        1.209       32.116      298.614
```

(**Fig 5)**

After the Regression Diagnosis, we get the equation of the Multiplier Linear Regression: RentPrice=872.388+1.209*Sqft+32.116*Beds+298.614*Baths

Lastly, we apply the predict function to rentprice model and new data in Fig 6. Based on the multiple linear regression model and the given parameters, we have the predicted price. Below are two examples with different square feet, number of bedrooms and bathrooms.

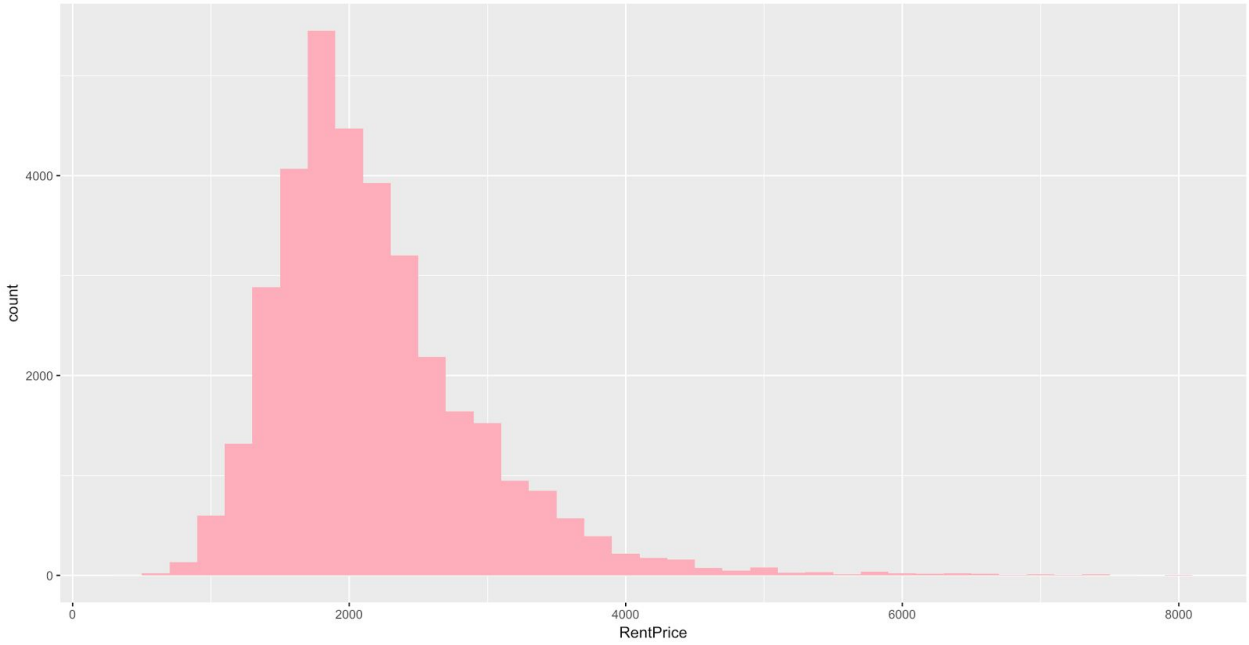
```
> model = lm(RentPrice~Sqft+Beds+Baths,data=CL_new)
> newdata = data.frame(Sqft=1450, Beds=2, Baths=2)
> predict(model, newdata)
       1
3287.334
```

(**Fig 6)**

```
> newdata1= data.frame(Sqft=1600, Beds=3, Baths=2)
> predict(model, newdata1)
       1
3500.845
```

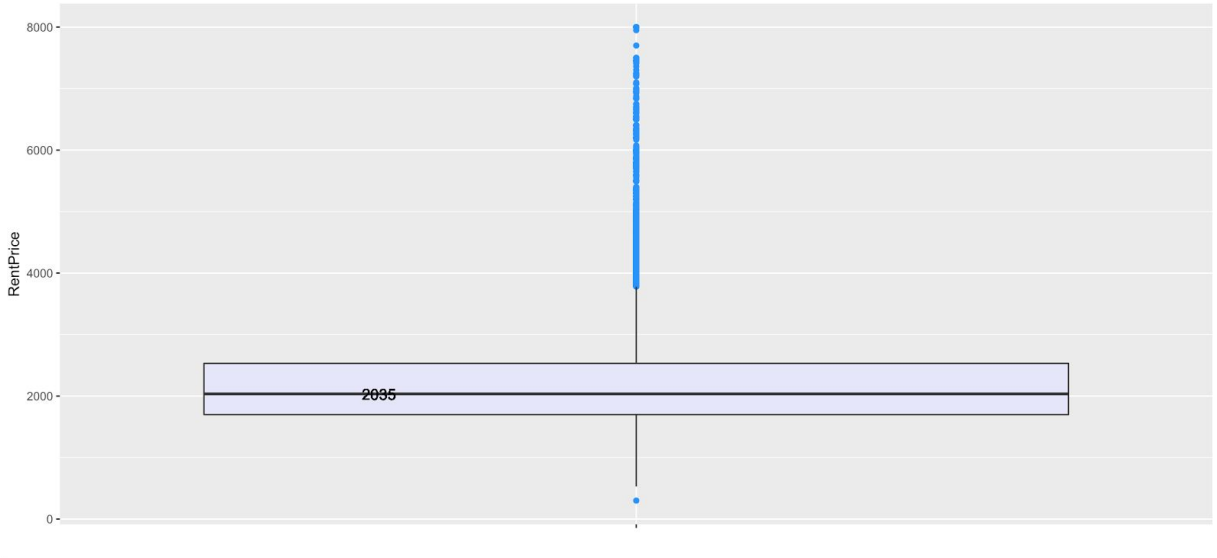
**R code for the project 2**

```
library(ggplot2)
library(dplyr)
Totaldata <- read.csv("~/Downloads/cleanCL.csv")
View(Totaldata)

#Column selection
vars <- c("listingTitle", "listingDate", "listingMonth", "cleanRent", "cleanBeds",
"cleanBaths", "cleanSqft", "matchAddress", "matchAddress2", "lat", "lng")
CLdata <- Totaldata[, vars]
View(CLdata)
str(CLdata)

#Change some column names to make more understandable and representative
colnames(CLdata) [4] <-"RentPrice"
colnames(CLdata) [5] <-"Beds"
colnames(CLdata) [6] <-"Baths"
colnames(CLdata) [7] <-"Sqft"
View(CLdata)

#Filter out data without NA
min(CLdata[,5])
which(is.na(CLdata$listingDate))
which(is.na(CLdata$Sqft))
CL_new<- CLdata[complete.cases(CLdata), ]
View(CL_new)

#The spread of rent price
#According to the graph of the rent price distribution, we can see that prices vary
drastically.
ggplot(data=CL_new ,aes(x= RentPrice))+geom_histogram(binwidth=200,
fill='lightpink1')
```



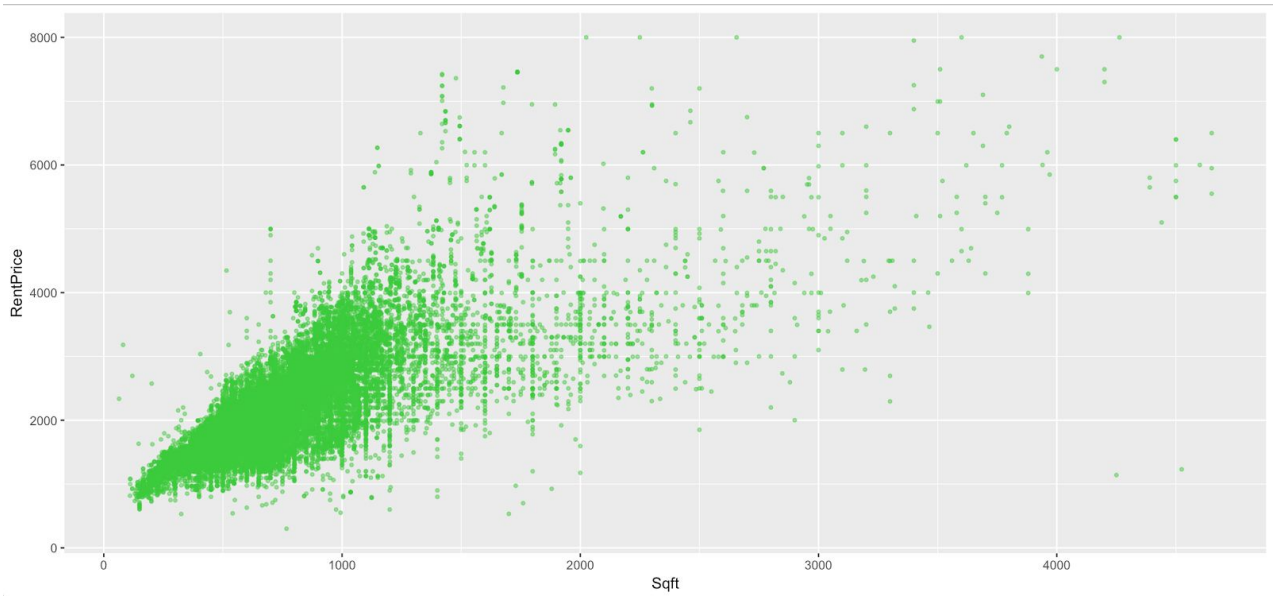

(**Fig 7)**

#Median rent price is $2035
ggplot(data=CL_new,aes(x='',y=RentPrice))+
  geom_boxplot(outlier.color='dodgerblue',fill='lavender')+
geom_text(aes(x='',y=median(CL_new$RentPrice),label=median(CL_new$RentPrice)),size=4,hjust=8)+xlab(label = '')



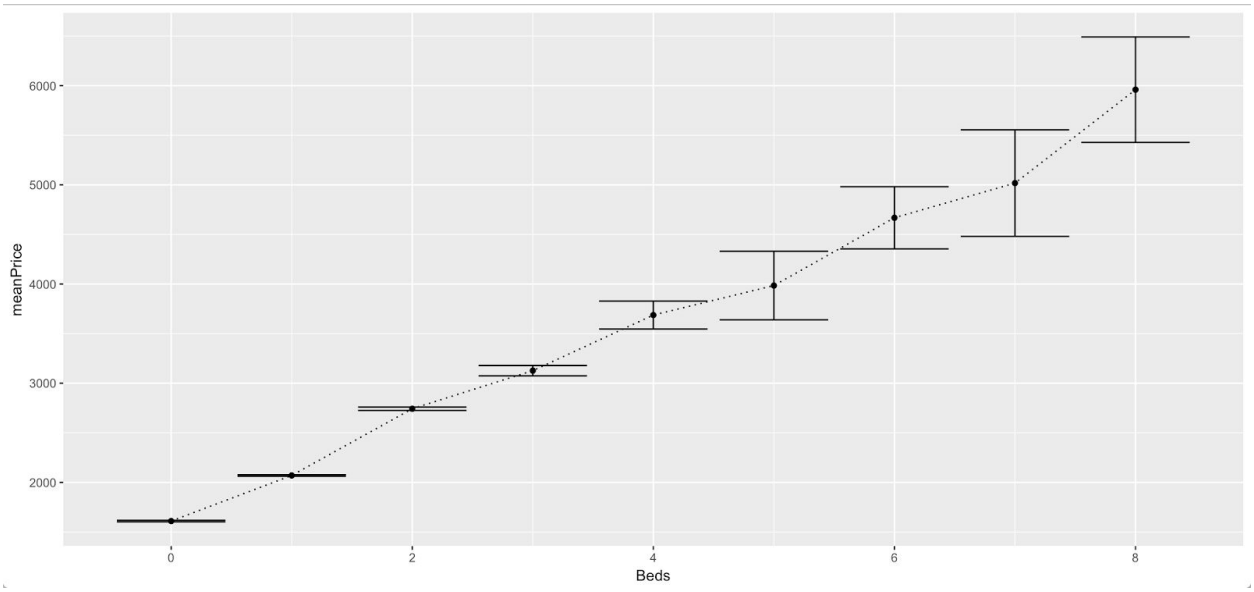

(**Fig 8)**
#We explored the relationship between Sqft and RentPrice

#Typical housing price with Sqft clusters around $2000 and 1000Sqft, as Sqft increases so does price
#The majority of sqft is 1000, and rent price $2000
```
ggplot(data=CL_new,aes(x=Sqft,y=RentPrice))+
  geom_point(alpha=0.5,size=0.9,color='limegreen')
```



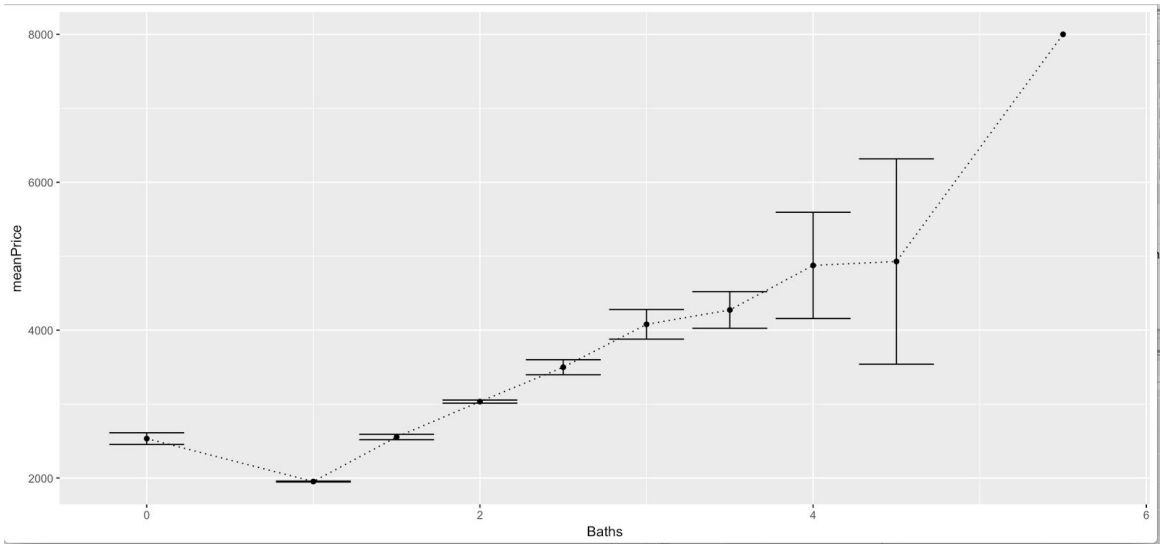

(**Fig 9)**

#Just as we thought: More bedrooms, higher price.
```
library(dplyr)
CL_new%>%
  group_by(Beds)%>%

summarize(meanPrice=mean(RentPrice),priceLow=mean(RentPrice)-1.96*sd(RentPrice)/sqrt(n()),priceHigh=mean(RentPrice)+1.96*sd(RentPrice)/sqrt(n()))%>%
  ungroup()%>%
  ggplot(aes(x=Beds,y=meanPrice))+
  geom_errorbar(aes(ymin=priceLow,ymax=priceHigh))+
  geom_line(aes(x=Beds,y=meanPrice,group=1),linetype=3)+
  geom_point(aes(x=Beds,y=meanPrice,group=1),size=1.5)
```




**(Fig 10)**

```
#Most as what we thought: More bathrooms, higher price.
CL_new%>%
  group_by(Baths)%>%

summarize(meanPrice=mean(RentPrice),priceLow=mean(RentPrice)-1.96*sd(RentPrice)/sqrt(n()),priceHigh=mean(RentPrice)+1.96*sd(RentPrice)/sqrt(n()))%>%
  ungroup()%>%
  ggplot(aes(x=Baths,y=meanPrice))+
  geom_errorbar(aes(ymin=priceLow,ymax=priceHigh))+
  geom_line(aes(x=Baths,y=meanPrice,group=1),linetype=3)+
  geom_point(aes(x=Baths,y=meanPrice,group=1),size=1.5)
```



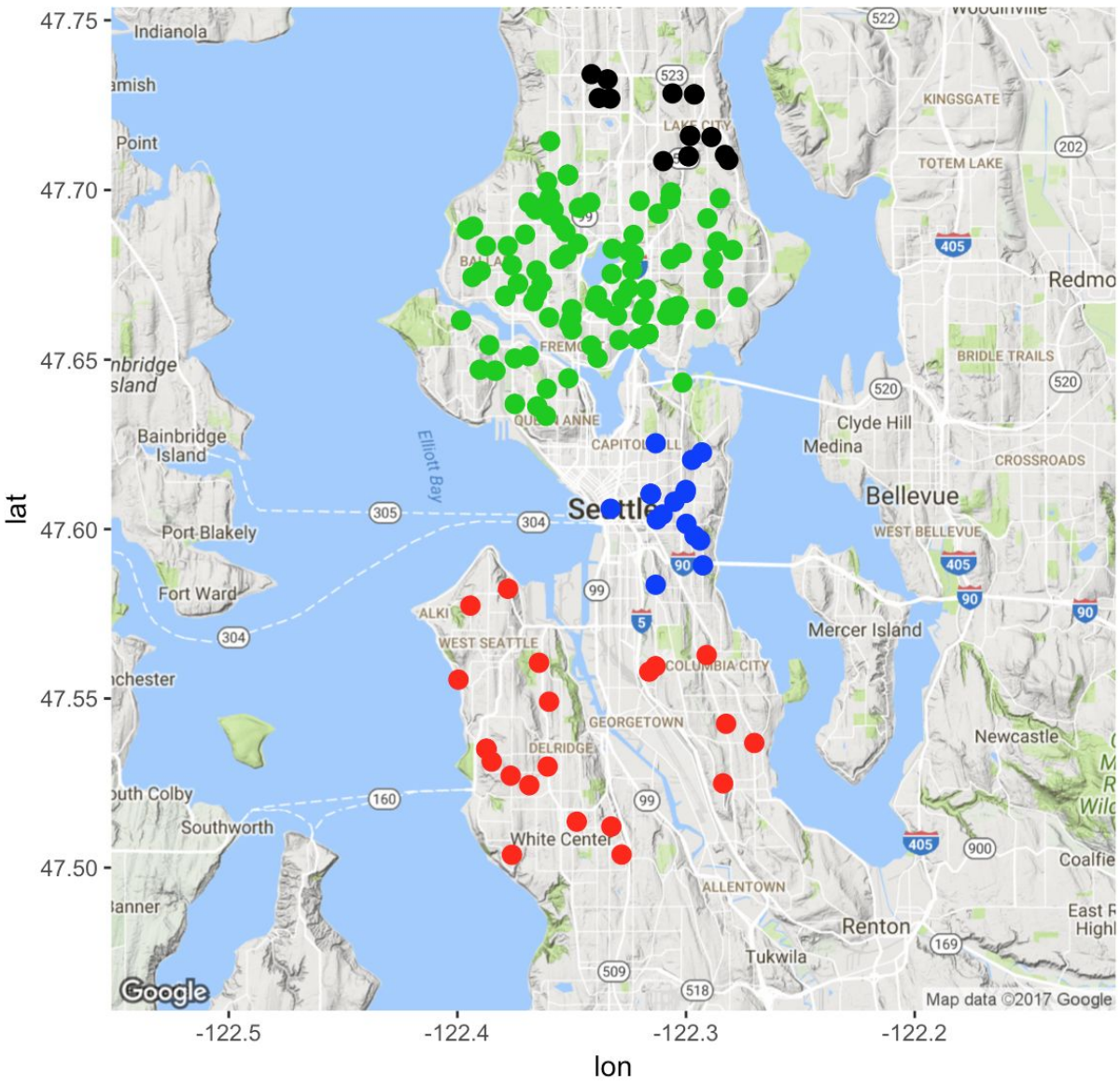

**(Fig 11)**

#However, there is one interesting thing catches our eyes when we try to plot the relationship between mean rent price and number of bathroom: There are some apartments have 0 bathroom with around $2500 average rent price. After filtering out those apartments with 0 bathroom, we finally realize that those are studios or rooftop deck as described in Listing title column.

```
#Filter out value=0 in the Baths column
#Turned out that those are studios, and many are in top floor (Rooftop deck)
bath0 <- dplyr::filter(CL_new, Baths == 0)
View(bath0)

## Linear Regression
names(CL_new)
model1 = lm(RentPrice~Sqft, CL_new)
summary(model1)

model2=lm(RentPrice~Beds, CL_new)
summary(model2)

model3 = lm(RentPrice~Baths, CL_new)
Summary(model3)
```

```
##Multiple Linear Regression

#Find correlations between factors. Display the correlation coefficient
library(corrplot)
factor_Corr <- cor(CL_new[,-c(1:2,8:9)])
corrplot(factor_Corr,method="number")

#Correlation Matrix and Probability value
library(psych)
corr.test(CL_new[,c(3:7,10:11)])

#Multiple Linear Regression

lm1 = lm(RentPrice~Sqft+Beds, data= subset(CL_new, matchAddress =="40th Ave
NE"))
summary(lm1)

model = lm(RentPrice~Sqft+Beds+Baths,data=CL_new)
summary(model)
anova(model)

#Regression Diagnostics
par(mfrow=c(2,2))
plot(model)

model
#Equation of Model: RentPrice=872.388+1.209*Sqft+32.116*Beds+298.614*Baths

#Wrap the prarameters inside a new data frame named newdata
newdata = data.frame(Sqft=1450, Beds=2, Baths=2)
newdata1= data.frame(Sqft=1600, Beds=3, Baths=2)

#Apply the predict function to rentprice lm and new data
predict(model, newdata)
predict(model, newdata1)

#Subset 3 Bedrooms & 1 Bathrooms apartemnts
threeb1b <- dplyr::filter(CL_new, Beds == 3 & Baths == 1)
View(threeb1b)
```

```
##Clustering

d = earth.dist(CL_new)
library(geosphere)
geo.dist = function(threeb1b) {
  require(geosphere)
  d <- function(i,z){         # z[11:10] contain long, lat
    dist <- rep(0,nrow(z))
    dist[i:nrow(z)] <- distHaversine(z[i:nrow(z),11:10],z[i,11:10])
    return(dist) }
  dm <- do.call(cbind,lapply(1:nrow(threeb1b),d,threeb1b))
  return(as.dist(dm))}

km <- kmeans(geo.dist(threeb1b),centers=4)  # k-means, 4 clusters
hc <- hclust(geo.dist(threeb1b))            # hierarchical clustering, dendrogram
clust <- cutree(hc, k=4)              # cut the dendrogram to generate 4 clusters

threeb1b$clust <- cutree(hc,k=4)
View(threeb1b)


# Map with 3B1B geographic location in Seattle
library(maps)
library(ggmap)
seattle = get_map(location="Seattle",zoom=11)
ggmap(seattle) +
  geom_point(data=threeb1b,aes(x=lng,y=lat),size=3,color=clust)
```




(**Fig 12)**

```
Residual standard error: 6.684 on 2 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:  0.9999
F-statistic: 2.29e+04 on 2 and 2 DF,  p-value: 4.366e-05
```

**(Fig13).** Less variability when restricting the model to specific localities. There are multiples of the same localities in the dataset, which is expected, as many different addresses can be on the same street, for example.