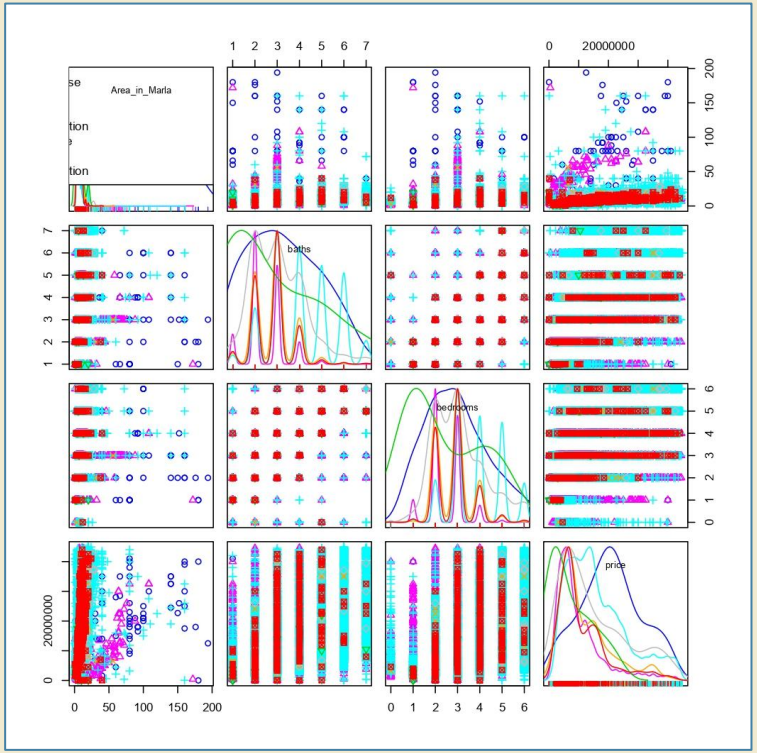
Pakistan
House Price Sales 2023
Analysis

Group 6 :

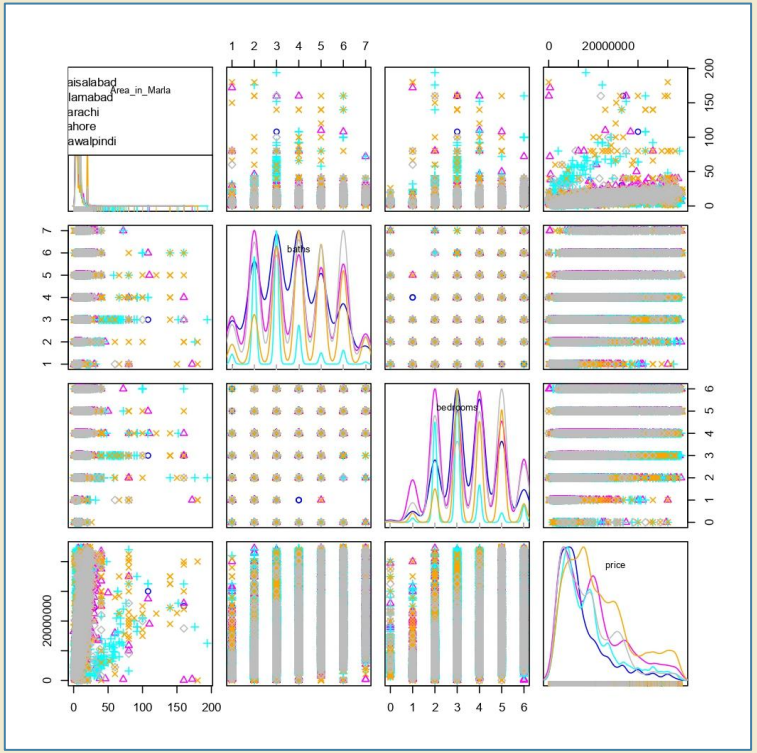
ChihHao Yuan

Ching Yu Hsu

1- Generate charts to display relationships



by property type



by city

2a)- Descriptive statistics of all quantitative variables

by property type

```
## max highest house price by property_type
Pakistan_House_Price_Sales_2023 %>%
  group_by(property_type) %>%
  summarize(max_price = max(price)) %>%
  arrange(desc(max_price))
```

property_type <chr>	max_price <dbl>
House	44900000
Flat	44500000
Upper Portion	44000000
Lower Portion	42500000
Penthouse	42500000
Farm House	41000000
Room	22000000

```
## min lowest house price by property_type
Pakistan_House_Price_Sales_2023 %>%
  group_by(property_type) %>%
  summarize(min_price = min(price)) %>%
  arrange(desc(min_price))
```

property_type <chr>	min_price <dbl>
Penthouse	1100000
Farm House	25000
Lower Portion	24000
Flat	23000
Upper Portion	23000
Room	20000
House	16000

```
Variable: price
                mean       sd      IQR       0%      25%      50%      75%     100%     n
Farm House  20708333 10307201 12000000    25000 14000000 21000000 26000000 41000000    75
Flat        10231007  8202029  8250000    23000  4750000  7500000 13000000 44500000 19383
House       16338953 10120352 13500000    16000  8500000 14000000 22000000 44900000 49022
Lower Portion 12167729  8853629 10000000    24000  6000000  8500000 16000000 42500000   660
Penthouse   14925864 11272432 16550000  1100000  5950000 11400000 22500000 42500000   184
Room         7295250  7370968  9175000    20000  2200000  3400000 11375000 22000000    12
Upper Portion 11339425  8457433  9500000    23000  5500000  8000000 15000000 44000000  1611
```

2b)- Descriptive statistics of all quantitative variables by city

```
## max highest house price by city
Pakistan_House_Price_Sales_2023 %>%
  group_by(city) %>%
  summarize(max_price = max(price)) %>%
  arrange(desc(max_price))
```

city <chr>	max_price <dbl>
Lahore	44900000
Karachi	44800000
Islamabad	44500000
Rawalpindi	44500000
Faisalabad	42500000

```
## min lowest house price by city
Pakistan_House_Price_Sales_2023 %>%
  group_by(city) %>%
  summarize(min_price = min(price)) %>%
  arrange(desc(min_price))
```

city <chr>	min_price <dbl>
Faisalabad	1200000
Karachi	25000
Rawalpindi	25000
Islamabad	23000
Lahore	16000

```
Variable: price
                mean       sd      IQR       0%      25%       50%       75%      100%     n
Faisalabad  10784274  8093139  8300000  1200000  5200000   8500000  13500000  42500000  1611
Islamabad   14626620 10204646 13300000    23000  6500000  13000000  19800000  44500000  8794
Karachi     12904060  9115364 10300000    25000  6200000  10800000  16500000  44800000 27210
Lahore      16780796 10655225 14000000    16000  8500000  14000000  22500000  44900000 26221
Rawalpindi  13054792  8904682 12000000    25000  6000000  11000000  18000000  44500000  7111
```

2c)- Correlation matrix of all quantitative variables

	price	bedrooms	baths	Area_in_Marla
price	1.0000000	0.5829138	0.6246022	0.5933768
bedrooms	0.5829138	1.0000000	0.8504592	0.3877019
baths	0.6246022	0.8504592	1.0000000	0.4204744
Area_in_Marla	0.5933768	0.3877019	0.4204744	1.0000000

The strongest correlation
→Baths and bedrooms(0.85)

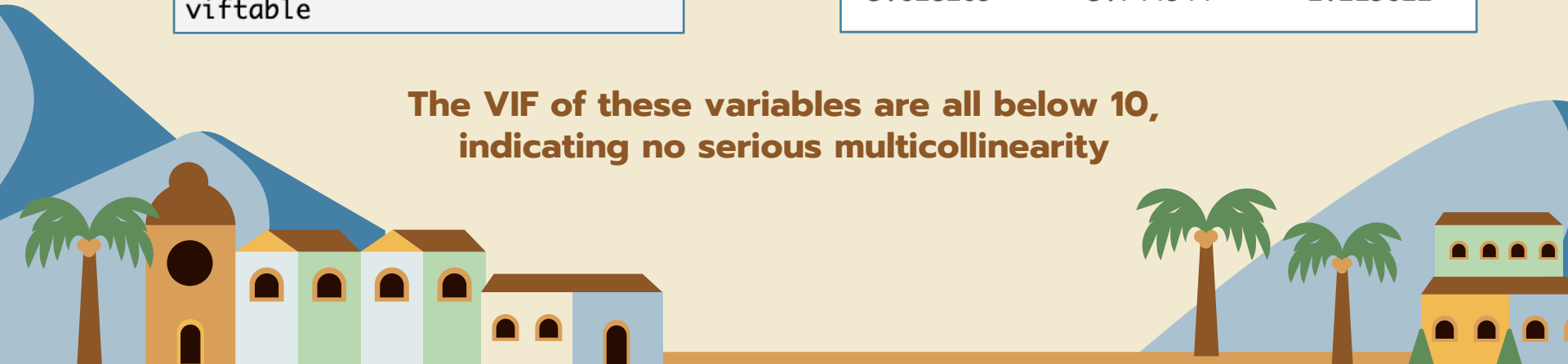
2d)- VIF of all quantitative variables

```
linearmodel <- lm(price~ bedrooms+baths+Area_in_Marla, data =
Pakistan_House_Price_Sales_2023)
```

```
library(car)
viftable <- vif(linearmodel)
viftable
```

```
bedrooms          baths Area_in_Marla
3.628209       3.744944       1.219622
```

The VIF of these variables are all below 10, indicating no serious multicollinearity



3a)- Convert property type and city to dummy variables

E	F	G	H	I	J
flat	house	lower_portion	penthouse	room	upper_portion
1	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
1	0	0	0	0	0
0	0	0	1	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0

N	O	P	Q
Islamabad	Karachi	Lahore	Rawalpindi
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	1	0

```
df_dum_all <- fastDummies::dummy_cols(
  df_hp_clean,
  select_columns = c("city", "property_type"),
  remove_first_dummy = TRUE
)
```

3b)- Create a new Y-variable (pricequartile)

A for highest quartile, B for second highest quartile, C for third highest quartile, D for lowest quartile of price

```
price_subset$pricequartile <- ifelse(price_subset$price > 20000000, "A",
                              ifelse(price_subset$price > 14517334, "B",
                              ifelse(price_subset$price > 6700000, "C", "D")))
```

Area_in_Marla	pricequartile
4.0	C
5.6	C
8.0	B
40.0	A
8.0	C
6.2	C
20.0	A
7.1	A
10.0	C
3.1	D
4.0	D
10.0	B

3c)- Run a linear regression for price as the Y-variable against property type, city, bedrooms, bathrooms, area as the X-variables.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -56549650     864288 -65.429   <2e-16 ***
flat         50036615     848323  58.983   <2e-16 ***
house        50292794     850322  59.146   <2e-16 ***
lower_portion 49238896    883787  55.714   <2e-16 ***
penthouse    49642583     971262  51.111   <2e-16 ***
room         47208510    2077094  22.728   <2e-16 ***
upper_portion 48938719    863121  56.700   <2e-16 ***
Islamabad     2914485     179562  16.231   <2e-16 ***
Karachi       4502810     173111  26.011   <2e-16 ***
Lahore        2842180     169429  16.775   <2e-16 ***
Rawalpindi     278787     181747   1.534    0.125
bedrooms      1063689      38386  27.711   <2e-16 ***
baths         2239058      32935  67.985   <2e-16 ***
Area_in_Marla  801231       5201 154.063   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6570000 on 70933 degrees of freedom
Multiple R-squared:  0.5666,    Adjusted R-squared:  0.5665
F-statistic:  7133 on 13 and 70933 DF,  p-value: < 2.2e-16
```

If there is a house located in Lahore, with 3 bedrooms and 2 baths and a size of 10 Marla

The predict price is

- 56549650 +50292794+2842180+3*1063689+2*2239058+10*801231

= $12,266,817

3d)- Run a SVM for pricequartile as the Y-variable against property type, city, bedrooms, bathrooms area as the X-variables.

Set the training and testing data

```r
set.seed(42)
ml_index <- sample(nrow(df_hp_final),
                   0.7 * nrow(df_hp_final),
                   replace = FALSE)

hp_train <- df_hp_final[ml_index,]
hp_test <- df_hp_final[-ml_index,]
```

Run the SVM model

```r
svm_model <- svm(pricequartile ~ bedrooms + baths + Area_in_Marla +
                   city_Islamabad +city_Karachi + city_Lahore + city_Rawalpindi +
                   property_type_Flat + property_type_House +
property_type_Lower_Portion + property_type_Penthouse + property_type_Room +
property_type_Upper_Portion,
                 data = hp_train)
```

Predict

```r
```{r}
svm_predict <- predict(svm_model, newdata = hp_test[,-15], type = "response")
```
```

SVM Model and Accuracy

```r
library(caret)
confusionMatrix(hp_test$pricequartile, svm_predict)
```

Accuracy : 0.6638

```
Confusion Matrix and Statistics

          Reference
Prediction    A    B    C    D
         A 3800  101 1171   26
         B 1163  246 1874   72
         C  313   70 5884 1279
         D   21    1 1064 4200

Overall Statistics

               Accuracy : 0.6638
                 95% CI : (0.6575, 0.6702)
    No Information Rate : 0.4695
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5237

 Mcnemar's Test P-Value : < 2.2e-16
```

3e)- Run a neural network with 5 hidden nodes for price as the Y-variable against property type, city, bedrooms, bathrooms, area as the X-variables.

Function definition

```
normalize <- function(x){return((x-min(x))/(max(x)-min(x)))}
denormalize <- function(y,x){return(y*(max(x)-min(x))+min(x))}
```

```
df_hp_final_norm <- as.data.frame(lapply(df_hp_final[,-15], normalize))
```

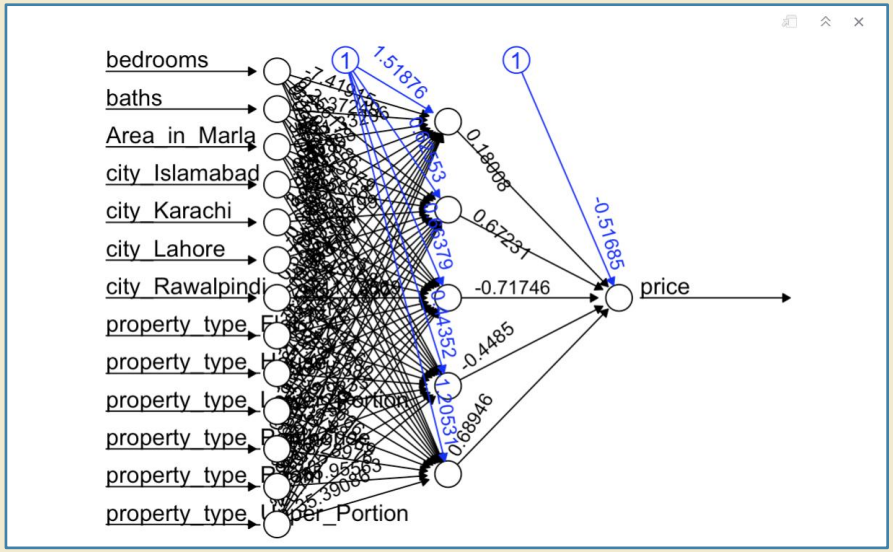
Split the dataset to training and testing

```
set.seed(42)
n_net_hp_train <- df_hp_final_norm[ml_index,]
n_net_hp_test <- df_hp_final_norm[-ml_index,]
```

Run the neural network model

```
set.seed(42)
n_net_model <- neuralnet(price ~ bedrooms + baths + Area_in_Marla +
                           city_Islamabad + city_Karachi + city_Lahore +
city_Rawalpindi +
                           property_type_Flat + property_type_House +
property_type_Lower_Portion + property_type_Penthouse + property_type_Room +
                           property_type_Upper_Portion,
                         hidden = 5,
                         data = n_net_hp_train,
                         lifesign = "minimal",
                         linear.output = TRUE,
                         threshold = 0.05)
```

Neural Network Model and Correlation




```
n_net_result <- compute(n_net_model, n_net_hp_test[,-1])
```

```
price_denorm <- denormalize(n_net_result$net.result,
df_hp_final_norm$price[n_net_index])
```

```
actual_price <- df_hp_final_norm$price[-n_net_index]
```

```
cor(price_denorm, actual_price)
```

Correlation

:0.84

3f)- Run a K-nearest neighbor analysis on pricequartile as the Y-variable against property type, city, bedrooms, bathrooms, area as the X-variables.

```
knn_hp_train_labels <- df_hp_final[ml_index, 15, drop = TRUE]
knn_hp_test_labels <- df_hp_final[-ml_index, 15, drop = TRUE]
```

Remove the price column

```
knn_hp_train <- n_net_hp_train[,-1]
knn_hp_test <- n_net_hp_test[,-1]
```

Error!!!!!

```
```{r}
knn_model <- class::knn(train = knn_hp_train,
                        test = knn_hp_test,
                        cl = knn_hp_train_labels,
                        k = 15)
```

Error in class::knn(train = knn_hp_train, test = knn_hp_test, cl = knn_hp_train_labels, :
  too many ties in knn
```

Use “FNN” to fix the problem of “tie”

```
install.packages("FNN")
library(FNN)

knn_model <- FNN::knn(train = knn_hp_train,
                      test = knn_hp_test,
                      cl = knn_hp_train_labels,
                      k = 15)
```

K-nearest neighbor Model and Accuracy

Total Observations in Table: 21285

knn_hp_test_labels	knn_model A	B	C	D	Row Total
A	4006	427	635	30	5098
	0.786	0.084	0.125	0.006	0.240
	0.722	0.231	0.073	0.006	
	0.188	0.020	0.030	0.001	
B	1126	840	1318	71	3355
	0.336	0.250	0.393	0.021	0.158
	0.203	0.454	0.151	0.014	
	0.053	0.039	0.062	0.003	
C	382	525	5657	982	7546
	0.051	0.070	0.750	0.130	0.355
	0.069	0.284	0.646	0.191	
	0.018	0.025	0.266	0.046	
D	31	59	1146	4050	5286
	0.006	0.011	0.217	0.766	0.248
	0.006	0.032	0.131	0.789	
	0.001	0.003	0.054	0.190	
Column Total	5545	1851	8756	5133	21285
	0.261	0.087	0.411	0.241	

Confusion Matrix and Statistics

Prediction	Reference A	B	C	D
A	4006	1126	382	31
B	427	840	525	59
C	635	1318	5657	1146
D	30	71	982	4050

Overall Statistics

Accuracy : 0.6837
95% CI : (0.6774, 0.69)
No Information Rate : 0.3545
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5596

Mcnemar's Test P-Value : < 2.2e-16

Accuracy : 0.6837

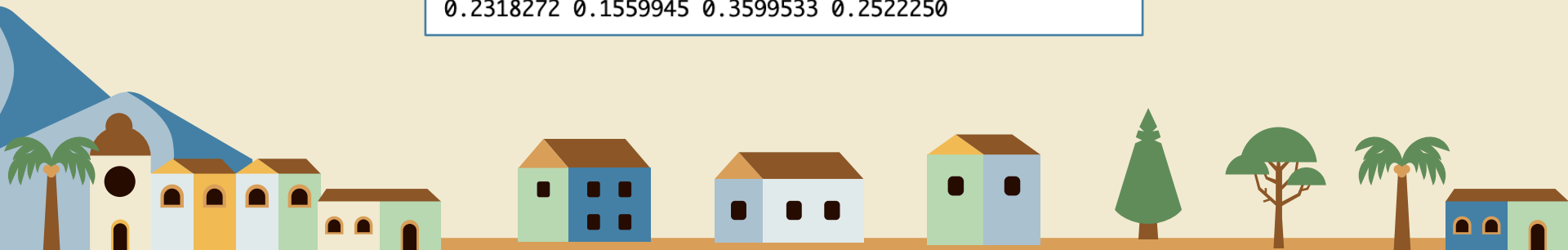
3g)- Run a Naïve Bayes analysis on pricequartile as the Y-variable against property type, city, bedrooms, bathrooms, area as the X variables

Run Naïve Bayes model

```
NB_model <- naiveBayes(pricequartile ~ bedrooms + baths + Area_in_Marla +
                  city_Islamabad +city_Karachi + city_Lahore + city_Rawalpindi +
                  property_type_Flat + property_type_House +
property_type_Lower_Portion + property_type_Penthouse + property_type_Room +
property_type_Upper_Portion,
                data = hp_train,
                laplace = 1)
```

```
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
         A          B          C          D
0.2318272 0.1559945 0.3599533 0.2522250
```



Naïve Bayes Model and Accuracy

```
Conditional probabilities:
   bedrooms
Y       [,1]     [,2]
  A 4.459220 1.0335370
  B 3.992384 1.0453536
  C 3.296095 1.0143520
  D 2.380329 0.9195205

   baths
Y       [,1]     [,2]
  A 4.924954 1.2680648
  B 4.325674 1.3159199
  C 3.458156 1.1454184
  D 2.259460 0.8637639

   Area_in_Marla
Y       [,1]     [,2]
  A 12.638965 7.419525
  B  8.629779 5.401589
  C  6.029900 3.353751
  D  3.928221 3.414935

   city_Islamabad
Y       [,1]     [,2]
  A 0.1283766 0.3345232
  B 0.1449593 0.3520826
  C 0.1067353 0.3087850
  D 0.1322050 0.3387270
```

```
   city_Karachi
Y       [,1]     [,2]
  A 0.2745592 0.4463113
  B 0.3596231 0.4799209
  C 0.4157530 0.4928651
  D 0.4517005 0.4976816

   city_Lahore
Y       [,1]     [,2]
  A 0.5001303 0.5000217
  B 0.3896992 0.4877135
  C 0.3510853 0.4773229
  D 0.2613763 0.4394020

   city_Rawalpindi
Y       [,1]     [,2]
  A 0.08659776 0.2812569
  B 0.08932490 0.2852305
  C 0.09946297 0.2992910
  D 0.12222577 0.3275595

   property_type_Flat
Y       [,1]     [,2]
  A 0.1183011 0.3229783
  B 0.1670324 0.3730289
  C 0.2717610 0.4448798
  D 0.4767683 0.4994799
```

```
   property_type_House
Y       [,1]     [,2]
  A 0.8574655 0.3496127
  B 0.7992771 0.4005670
  C 0.6931081 0.4612171
  D 0.4744531 0.4993669

   property_type_Lower_Portion
Y         [,1]      [,2]
  A 0.006427517 0.07991720
  B 0.008519427 0.09191266
  C 0.008670844 0.09271538
  D 0.012773431 0.11229994

   property_type_Penthouse
Y         [,1]      [,2]
  A 0.003213758 0.05660131
  B 0.001678069 0.04093250
  C 0.002125755 0.04605817
  D 0.003432860 0.05849229

   property_type_Room
Y          [,1]       [,2]
  A 8.685833e-05 0.009319782
  B 1.290822e-04 0.011361436
  C 1.118819e-04 0.010577126
  D 5.588376e-04 0.023634084

   property_type_Upper_Portion
Y         [,1]      [,2]
  A 0.01216017 0.1096054
  B 0.02194398 0.1465101
  C 0.02355113 0.1516501
  D 0.03153441 0.1747639
```

```
Confusion Matrix and Statistics

          Reference
Prediction    A    B    C    D
         A 3741 1696 1595  214
         B  111  132  146   45
         C 1122 1336 4558 2047
         D  124  191 1247 2980

Overall Statistics

               Accuracy : 0.5361
                 95% CI : (0.5294, 0.5428)
    No Information Rate : 0.3545
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3478

 Mcnemar's Test P-Value : < 2.2e-16
```

Accuracy : 0.5361

3h)- Summarize the accuracy of all models

Model	Accuracy/Correlation	P-value
SVM	0.6638	2.2e-16
Neural Network	0.84 (correlation)	
K-nearest neighbor	0.6837	2.2e-16
Naïve Bayes	0.5361	2.2e-16

4)- List of Lessons Learned

a. <u>Which technique worked best (accuracy)?</u>

- It’s not ideal to compare all models directly, because Neural Network is evaluated by correlation, but others are evaluated by accuracy.
- This neural network has a strong predictive (correlation = 0.84)
- If we just consider the accuracy, K-nearest neighbor works best (accuracy=0.6837)

4)- List of Lessons Learned

b. <u>What insights do you have on which variables influence housing prices in Pakistan?</u>

- Based on the linear regression, we found the p-value of property type, city, bedrooms, bathrooms, area are statistically significant.
- So, we can conclude these 5 variables will impact house prices in Pakistan.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -56549650     864288 -65.429   <2e-16 ***
flat         50036615     848323  58.983   <2e-16 ***
house        50292794     850322  59.146   <2e-16 ***
lower_portion 49238896    883787  55.714   <2e-16 ***
penthouse    49642583     971262  51.111   <2e-16 ***
room         47208510    2077094  22.728   <2e-16 ***
upper_portion 48938719    863121  56.700   <2e-16 ***
Islamabad     2914485     179562  16.231   <2e-16 ***
Karachi       4502810     173111  26.011   <2e-16 ***
Lahore        2842180     169429  16.775   <2e-16 ***
Rawalpindi     278787     181747   1.534    0.125
bedrooms      1063689      38386  27.711   <2e-16 ***
baths         2239058      32935  67.985   <2e-16 ***
Area_in_Marla  801231       5201 154.063   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6570000 on 70933 degrees of freedom
Multiple R-squared:  0.5666,    Adjusted R-squared:  0.5665
F-statistic:  7133 on 13 and 70933 DF,  p-value: < 2.2e-16
```

Property type

city

bedrooms

bathrooms

area

Thanks!

Pakistan House Price Sales 2023 Analysis

Group 6:
Chih Hao Yuan, Ching Yu Hsu