Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos

Jorge García, 1220019
Tito Fajardo, 1201619

Proyecto 2

Dataset

Para el dataset de entrenamiento y testing se utilizó house_rent_result90, el cual se dividió en un **80%** para el entrenamiento del modelo y un **20%** para el testeo del mismo.

Description: df [6 × 13]

	X <int>	Posted.... <chr>	BHK <int>	Rent <int>	Size <int>	Floor <chr>	Area.Ty... <chr>
1	1	2022-0...	2	10000	1100	Ground ...	Super A...
2	2	2022-0...	2	20000	800	1 out of 3	Super A...
3	3	2022-0...	2	17000	1000	1 out of 3	Super A...
4	4	2022-0...	2	10000	800	1 out of 2	Super A...
5	5	2022-0...	2	7500	850	1 out of 2	Carpet ...
6	6	2022-0...	2	7000	600	Ground ...	Super A...

6 rows | 1-8 of 13 columns

Feature Engineering

1. Limpieza de Campos Nulos

Para que nuestros datos fueran correctos y precisos se inició limpiando los datos nulls de cada una de las columnas del dataset las cuales tuvieran dichos datos.

Limpieza de nulls

```
{r}
colsNulls <-colnames(Odataset)[!complete.cases(t(Odataset))]
colsNulls
```

```
character(0)
```

2. Cambios de datos categóricos

Se agrupará por categoría y se obtendrá el promedio de la renta por categoría.

- Furnishing.Status

Furnishing.Status <chr>	MeanFurnish <dbl>
Furnished	21274.797
Semi-Furnished	3131.599
Unfurnished	-11813.013

3 rows

- Floor

A tibble: 460 × 2

Floor <chr>	MeanFloor <dbl>
24 out of 24	700000.00
7 out of 20	600000.00
36 out of 81	380000.00
18 out of 20	353500.00
19 out of 85	350000.00
39 out of 60	350000.00
45 out of 60	350000.00
20 out of 41	330000.00
8 out of 27	330000.00
19 out of 33	310000.00

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos

Jorge García, 1220019
Tito Fajardo, 1201619

- Area.Type

A tibble: 3 × 2

Area.Type <chr>	MeanAreaT <dbl>
Carpet Area	51414.02
Super Area	18436.71
Built Area	10500.00

- Area.Locality

A tibble: 2,098 × 2

Area.Locality <chr>	MeanAreaL <dbl>
Marathahalli	715780.00
Lady Ratan Tower, Worli	700000.00
Bandra East	600000.00
Vettuvankeni	600000.00
Altamount Road	500000.00
Rustomjee Elements, Andheri West	400000.00
World One Tower Mumbai, Worli	380000.00
Deonar	350000.00
Green Park	350000.00
Indiabulls Blu, Worli Naka Acharaya Atre Chowk	350000.00
Lodha World Crest, Lower Parel	350000.00
Sundar Nagar	350000.00
Anand Niketan	330000.00

- City

A tibble: 6 × 2

City <chr>	MeanCity <dbl>
Mumbai	81103.28
Delhi	30230.35
Bangalore	25593.93
Chennai	21568.14
Hyderabad	20442.06
Kolkata	11599.94

6 rows

- Tenant.Preferred

A tibble: 3 × 2

Tenant.Preferred <chr>	MeanTenant <dbl>
Family	46744.58
Bachelors	43689.01
Bachelors/Family	30616.38

3 rows

- Point.of.Contact

A tibble: 3 × 2

Point.of.Contact <chr>	MeanPoC <dbl>
Contact Agent	72098.91
Contact Owner	16630.84
Contact Builder	5500.00

3 rows

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos

Jorge García, 1220019
Tito Fajardo, 1201619

3. Cambio de columnas (Datos calculados)

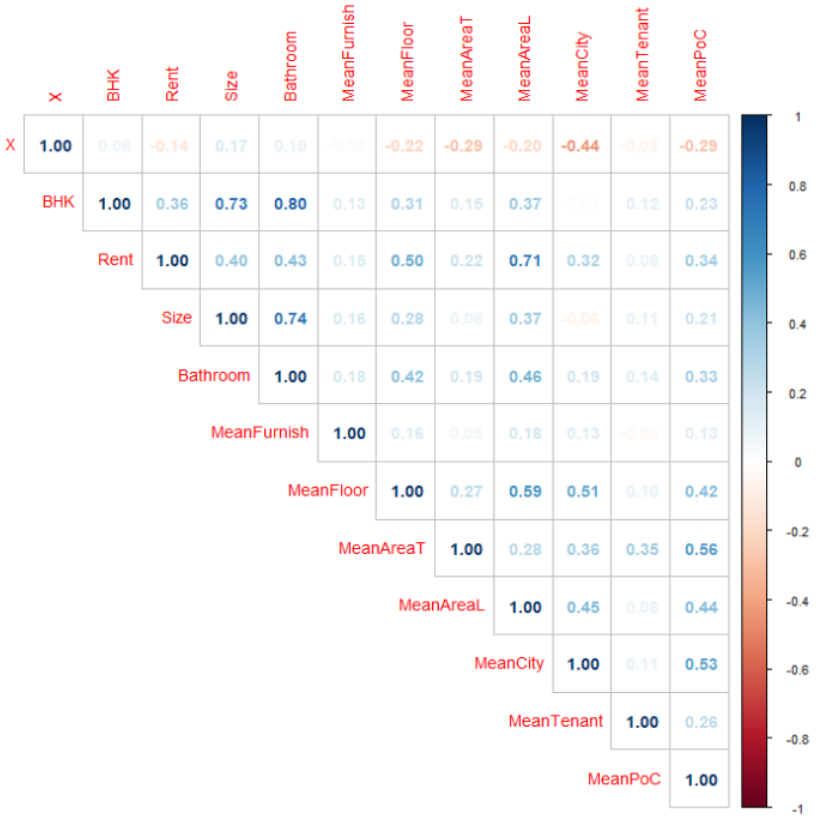
Se remplazan las columnas del dataset original por las columnas con las medias calculadas.

Description: df [4,278 × 12]

X <int>	BHK <int>	Rent <int>	Size <int>	Bathroom <int>	MeanFurnish <dbl>	MeanFloor <dbl>	MeanAreaT <dbl>	MeanAreaL <dbl>	MeanCity <dbl>	MeanTenant <dbl>	MeanPoC <dbl>
2803	1	8500	550	1	-11813.013	17742.20	51414.02	17550.000	30230.35	43689.01	72098.91
597	3	75000	1150	3	-11813.013	75000.00	51414.02	72500.000	81103.28	30616.38	72098.91
4744	3	35000	1750	3	3131.599	20684.17	51414.02	29428.571	20442.06	30616.38	72098.91
1805	1	10000	650	1	3131.599	75831.88	51414.02	19250.000	25593.93	43689.01	72098.91
4369	3	95000	2700	3	21274.797	95000.00	51414.02	45019.231	20442.06	43689.01	72098.91
950	2	40000	755	2	3131.599	40000.00	51414.02	40000.000	81103.28	43689.01	72098.91
1219	1	21000	450	2	-11813.013	40500.00	51414.02	21000.000	81103.28	43689.01	72098.91
674	3	160000	1050	3	21274.797	86000.00	51414.02	160000.000	81103.28	43689.01	72098.91
476	2	7000	600	1	-11813.013	14169.31	18436.71	6200.000	11599.94	43689.01	72098.91
4003	3	49000	1800	3	-11813.013	64500.00	18436.71	59500.000	20442.06	43689.01	72098.91
1691	1	10000	415	1	3131.599	26060.44	51414.02	10000.000	25593.93	30616.38	72098.91
73	2	16000	850	1	3131.599	14204.06	51414.02	22857.143	11599.94	43689.01	72098.91

Correlación de los datos

La variable con la mayor correlación con la variable **Rent** es: **MeanAreaL**



Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos

Jorge García, 1220019
Tito Fajardo, 1201619

Pruebas y Experimentos realizados

- Experimento 1

```r
{r}
experimento1 <- dstrain %>%
  lm(formula = Rent~MeanAreaL)# Pruebas, realizar operaciones entre las columnas
del dataset

yhat <- predict(experimento1, dstest)

rmseExp1 <- rmse(yhat,dstest$Rent)
rmseExp1
EXP1<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP1
```

RMSE = 19790.35

- Experimento 2

Se agrega la columna MeanFloor

```r
experimento2 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor)# Pruebas, realizar operaciones entre las
columnas del dataset

yhat <- predict(experimento2, dstest)

rmseExp2 <- rmse(yhat,dstest$Rent)
rmseExp2
EXP2<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP2
```

RMSE = 19210.67

- Experimento 3

Se le agregan todas las columnas

```r
experimento3 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+BHK+Size+Bathroom+MeanAreaT
+MeanCity+MeanTenant+MeanPoC)# Pruebas, realizar operaciones entre las columnas del
dataset

yhat <- predict(experimento3, dstest)

rmseExp3 <- rmse(yhat,dstest$Rent)
rmseExp3
EXP3<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP3
```

RMSE = 17370.65

- Experimento 4

Se le quita la columna MeanPoC

```r
experimento4 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+BHK+Size+Bathroom+MeanAreaT
+MeanCity+MeanTenant)# Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento4, dstest)

rmseExp4 <- rmse(yhat,dstest$Rent)
rmseExp4
EXP4<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP4
```

RMSE = 17376.38

- Experimento 5

Se le quita la columna MeanTenant

Jorge García, 1220019
Tito Fajardo, 1201619

```
experimento5 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+BHK+Size+Bathroom+MeanAreaT
+MeanCity)# Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento5, dstest)

rmseExp5 <- rmse(yhat,dstest$Rent)
rmseExp5
EXP5<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP5
```

RMSE = 17345.21

- Experimento 6

  Se le agrega la columna MeanPoc

```
experimento6 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+BHK+Size+Bathroom+MeanAreaT
+MeanCity+MeanPoC)# Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento6, dstest)

rmseExp6 <- rmse(yhat,dstest$Rent)
rmseExp6
EXP6<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP6
```

RMSE = 17339.99

- Experimento 7

  Se elimina MeanCity

```
experimento7 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+BHK+Size+Bathroom+MeanAreaT
+MeanPoC)# Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento7, dstest)

rmseExp7 <- rmse(yhat,dstest$Rent)
rmseExp7
EXP7<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP7
```

RMSE = 17329.79

- Experimento 8

  Se elimina MeanAreaT

```
experimento8 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+BHK+Size+Bathroom+MeanPoC)#
Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento8, dstest)

rmseExp8 <- rmse(yhat,dstest$Rent)
rmseExp8
EXP8<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP8
```

RMSE = 17339.58

- Experimento 9

Jorge García, 1220019
Tito Fajardo, 1201619

Se agrega MeanAreaT y se elimina Bathroom

```
experimento9 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+BHK+Size+MeanAreaT+MeanPoC)#
Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento9, dstest)

rmseExp9 <- rmse(yhat,dstest$Rent)
rmseExp9
EXP9<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP9
```

RMSE = 17354.42

- Experimento 10

  Se agrega Bathroom y se elimina Size

```
experimento10 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+BHK+Bathroom+MeanAreaT+MeanPoC
)# Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento10, dstest)

rmseExp10 <- rmse(yhat,dstest$Rent)
rmseExp10
EXP10<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP10
```

RMSE = 17403.06

- Experimento 11

  Se agrega Size y se elimina BHK

```
experimento11 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+Size+Bathroom+MeanAreaT+MeanPoC
)# Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento11, dstest)

rmseExp11 <- rmse(yhat,dstest$Rent)
rmseExp11
EXP11<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP11
```

RMSE = 17321.37

- Experimento 12

  Se elimina MeanFurnish

```
experimento12 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+Size+Bathroom+MeanAreaT+MeanPoC)# Pruebas,
realizar operaciones entre las columnas del dataset

yhat <- predict(experimento12, dstest)

rmseExp12 <- rmse(yhat,dstest$Rent)
rmseExp12
EXP12<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP12
```

RMSE = 17324.27

- Experimento 13

  Se agrega MeanFurnish y se elimina MeanFloor

```
experimento13 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFurnish+Size+Bathroom+MeanAreaT+MeanPoC)# Pruebas
, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento13, dstest)

rmseExp13 <- rmse(yhat,dstest$Rent)
rmseExp13
EXP13<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP13
```

Jorge García, 1220019
Tito Fajardo, 1201619

RMSE = 17847.21

- Experimento 14

  Se agrega MeanFloor y se elimina MeanAreaL

```
experimento14 <- dstrain %>%
  lm(formula = Rent~MeanFloor+MeanFurnish+Size+Bathroom+MeanAreaT+MeanPoC)# Pruebas
, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento14, dstest)

rmseExp14 <- rmse(yhat,dstest$Rent)
rmseExp14
EXP14<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP14
```

RMSE = 31151.22

- Experimento 15

  Se cambia Bathroom por MeanBathroom

```
experimento15 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+Size+MeanBathroom+MeanAreaT
+MeanPoC)# Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento15, dstest)

rmseExp15 <- rmse(yhat,dstest$Rent)
rmseExp15
EXP15<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP15
```

RMSE = 29654.54

- Experimento 16

```
experimento16 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+Size+Bathroom+MeanAreaT)#
Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento16, dstest)

rmseExp16 <- rmse(yhat,dstest$Rent)
rmseExp16
EXP16<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP16
```

RMSE = 17318.53

- Experimento 17

  Debido a que se existen categorías en el archivo testing (el de 10%) que no estaban en el archivo de pruebas, se le agrego la media de la categoría para reemplazar los nulls que se agregaron.

```
#Experimento 17 - Se cambiaron los nulls de las categorias inexistentes del archivo testing por el promedio de las medias de los de esa categoría
```{r}
experimento17 <- dstrain %>%
  lm(formula = Rent~MeanAreaL+MeanFloor+MeanFurnish+Size+Bathroom+MeanAreaT)# Pruebas, realizar operaciones entre las columnas del dataset

yhat <- predict(experimento17, dstest)

rmseExp17 <- rmse(yhat,dstest$Rent)
rmseExp17
EXP17<- data.frame("Rownum"= dstest$X, "RentP" = yhat, "Rent"=dstest$Rent)
EXP17
```
```


```
[1] 17318.53
```

RMSE = 17.318.53

Jorge García, 1220019
Tito Fajardo, 1201619

```{r}
datasetPredict$MeanFurnish[is.na(datasetPredict$MeanFurnish)] <- FurnM
datasetPredict$MeanFloor[is.na(datasetPredict$MeanFloor)] <- FloorM
datasetPredict$MeanAreaT[is.na(datasetPredict$MeanAreaT)] <- AreaTM
datasetPredict$MeanAreaL[is.na(datasetPredict$MeanAreaL)] <- AreaLM
datasetPredict$MeanCity[is.na(datasetPredict$MeanCity)] <- CityM
datasetPredict$MeanTenant[is.na(datasetPredict$MeanTenant)] <- TenM
datasetPredict$MeanPoC[is.na(datasetPredict$MeanPoC)] <- PoCM
datasetPredict
```

Repositorio de Github: https://github.com/tjfv02/Proyecto2_Analisis-de-Datos.git