# Property Analysis

Caio di Felice Cunha

## Definition of the Business Problem

Our job will be to check if two of the variables (both categorical) are related and impact the general information available in the dataset. We will apply a statistical test indicated for categorical variables, the Chi-Square Test.

For this work we will be using a dummy dataset.

## Stage 1 - Collecting the Data and Unsdertand the data

```
# Loading the dataset
df = read.csv("data.csv")

# Visualizing the data
head(df)
```

```
##    Price  Size Zip_Code       Type_Property Status_Property Status_Rent
## 1 899950 37505  NW3 1RX           Apartment             Old      Rented
## 2 330000 37475   W3 6DR House With Backyard             Old       Empty
## 3 230000 37270  SW6 2RX           Apartment             Old      Rented
## 4 178000 37596  CR0 9LQ Penthouse Apartment             Old       Empty
## 5 180000 37396 SE27 9AW House With Backyard             Old       Empty
## 6 130000 37293 SW15 1HJ           Apartment             Old      Rented
##        City
## 1     Natal
## 2     Natal
## 3     Natal
## 4 Fortaleza
## 5     Natal
## 6     Natal
```

```
# Separating x and y
x <- df$Type_Property
y <- df$Status_Property
```

```
# values cross table
table(x,y)
```

```
##                           y
## x                           New  Old
##   Apartment                 990 2901
##   House With Backyard       7  357
##   House without Backyard   19  961
##   Other                    14  656
##   Penthouse Apartment      43 1752
```

```
# percentage cross table
round(prop.table(table(x,y))*100,2)
```

```
##                         y
## x                          New   Old
##   Apartment              12.86 37.68
##   House With Backyard     0.09  4.64
##   House without Backyard  0.25 12.48
##   Other                   0.18  8.52
##   Penthouse Apartment     0.56 22.75
```

## Stage 2 - Defining the hypotheses:

- H0 = There is no relationship between x and y
- H1 = x and y are related

If the p-value is less than 0.05 we reject the H0

```
chisq.test(x,y)
```

```
##
##  Pearson’s Chi-squared test
##
## data:  x and y
## X-squared = 868.75, df = 4, p-value < 2.2e-16
```

## Stage 3 - Question

If we do not consider Apartment type properties, is there a difference in the test result?

```
## Way of Gaius

## Create the tables filtering to remove the "apartment"
x <- df$Type_Property[df$Type_Property != "Apartment"]
y <- df$Status_Property[df$Type_Property != "Apartment"]

# Making cross tables
table(x,y)
```

```
##                         y
## x                         New  Old
##   House With Backyard       7  357
##   House without Backyard   19  961
##   Other                    14  656
##   Penthouse Apartment      43 1752
```

```
round(prop.table(table(x,y))*100,2)
```

```
##                          y
## x                           New   Old
##   House With Backyard     0.18  9.37
##   House without Backyard  0.50 25.23
##   Other                   0.37 17.22
##   Penthouse Apartment     1.13 46.00
```

```
# Chi-Square test with the new database
chisq.test(x,y)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 0.79718, df = 3, p-value = 0.8501
```

```
chisq.test(table(x,y))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(x, y)
## X-squared = 0.79718, df = 3, p-value = 0.8501
```

## Disclaimer:

Disclaimer: a good part of this project was largely done in the Data Science Academy, Big Data Analytics with R and Microsoft Azure Machine Learning course (part of the Data Scientist training)

## End