DREAM HOME

Group project with R

Skills:
Programming with
Advanced Computer Languages

Theresa Maria Seehofer 20-600-375

Julian Thomas Albrecht 20-601-696

University of St. Gallen

December 18, 2020

Mario Silic

Contents

1. Introduction

House Purchase

You always wanted to buy your dream home in King County in the U.S state of Washington?

But you do not know how expensive it is?

House Sale

You want to sell your home to discover another place?

But you do not know the value of your home?

House Price Calculator

In this case, our real estate calculator helps you as a buyer or seller!

Just enter a few house features and our program calculates the appropriate price for you!

2. Requirements

The following program works with R/RStudio.

In order to run it, the following packages need to be installed:

- Openxlsx
- Leaflet
- Leaflet.extras

3. Program Structure

This program is a tool for potential buyers or sellers of a property. It asks the user for house features as input and calculates an estimated property value using a real-world dataset. The dataset contains house sale prices for King County, Washington from the time period May 2014 – May 2015. After importing the data, the program explores the dataset by estimating different variables and observations, so the user can become familiar with it. Different implemented visualizations illustrate for example the density of the properties or the price range in King County. Furthermore, the program estimates the effect of different house features on the house price with the aid of regression models. Finally, the user can enter his favored house features, which include among others the number of bedrooms, bathrooms and floors, to get an estimated property value as output.

Summarized, the code is structured in the following five parts:

1. **Load Packages and Import Data**
   Load packages that are required to import, process and visualize the data
2. **Explore the dataset and its variables**
   Analyze and adjust existing variables, create new variables
3. **Visualize Data**
   Visualize the location of the properties and their prices
4. **Regressions**
   Estimation of various regression models which will be the basis for the house price calculator
5. **House Price Calculator**: How much is my (future) house worth?

4. Overview of the Data

Variable	Description
Id	Unique ID for each home sold
Date	Date of home sale
Price	Price of each home sold
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms, where 0.5 accounts for a room with a toilet but no shower
Sqft_living / Sqm_living	Square feet / square meters of the apartment's interior living space
Sqft_lot / Sqm_lot	Square feet / square meters of the land space
Floors	Number of floors
Waterfront	A dummy variable that indicates whether the apartment is located at the waterfront or not
View	An index from 0 to 4 of how good the view of property is
Condition	An index from 1 to 5 on the condition of the apartment
Grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
Sqft_above / Sqm_above	Square feet / square meters of the interior housing space that is above ground level
Sqft_basement / Sqm_basement	Square feet / square meters of the interior housing space that is below ground level
Yr_build	The year the house was initially built
Yr_renovated	The year of the house's last renovation
Zipcode	Zipcode area where the house is located
Lat	Latitude
Long	Longitude
Sqft_living15 / Sqm_living15	Square feet / square meters of interior housing living space for the nearest 15 neighbors
Sqft_lot15 / Sqm_lot15	Square feet / square meters of the land space of the nearest 15 neighbors
Age	Age of the property
Age_r	Numbers of years since last renovation

5. Code with Output

This is an excerpt from the provided HTML file. Please also download the HTML file in order to be able to use the maps!

Dream Home: Your House Price Calculator

Julian Albrecht, Theresa Maria Seehofer

16 12 2020

Part 1: Load packages that are required to import, process and visualize the data

Dataset: House Prices from King County, USA (Source: https://www.kaggle.com/harlfoxem/housesalesprediction)

Load Packages:

```
library(openxlsx)
library(leaflet)
library(leaflet.extras)
```

Import Data:

```
kc_data <- read.xlsx("kc_house_data.xlsx")
```

Adjust the output format:

```
options(scipen = 6)
```

Part 2: Explore the dataset and its variables

Explore Data:

```
dim(kc_data)
```

```
## [1] 21613    21
```

The dataset includes 21 variables and 21613 observations (properties).

```
head(kc_data)
```

```
##           id            date   price bedrooms bathrooms sqft_living sqft_lot
## 1 7129300520 20141013T000000  221900        3      1.00        1180     5650
## 2 6414100192 20141209T000000  538000        3      2.25        2570     7242
## 3 5631500400 20150225T000000  180000        2      1.00         770    10000
## 4 2487200875 20141209T000000  604000        4      3.00        1960     5000
## 5 1954400510 20150218T000000  510000        3      2.00        1680     8080
## 6 7237550310 20140512T000000 1225000        4      4.50        5420   101930
##   floors waterfront view condition grade sqft_above sqft_basement yr_built
## 1      1          0    0         3     7       1180             0     1955
## 2      2          0    0         3     7       2170           400     1951
## 3      1          0    0         3     6        770             0     1933
## 4      1          0    0         5     7       1050           910     1965
## 5      1          0    0         3     8       1680             0     1987
## 6      1          0    0         3    11       3890          1530     2001
##   yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1            0   98178 47.5112 -122.257          1340       5650
## 2         1991   98125 47.7210 -122.319          1690       7639
## 3            0   98028 47.7379 -122.233          2720       8062
## 4            0   98136 47.5208 -122.393          1360       5000
## 5            0   98074 47.6168 -122.045          1800       7503
## 6            0   98053 47.6561 -122.005          4760     101930
```

The variables are typical house features like price, number of bedrooms, number of bathrooms, location etc. A detailed description of all variables can be found in the introduction file.

We do not need the House ID and the transaction date so they are removed:

```
kc_data <- kc_data[,3:21]
```

In the next step, we estimate two new variables that are important for the estimation of the property value. We use the existing variable year_built to calculate the age of the property and the variable year_renovated to estimate the number of years since the last renovation.

New variable 1: Age of the property

```
max(kc_data$yr_built)
```

```
## [1] 2015
```

```
kc_data$age <- 2015 - kc_data$yr_built
```

New variable 2: Number of years since last renovation

```
for(i in 1:length(kc_data$yr_renovated)) {
  if(kc_data$yr_renovated[i]==0) {
  kc_data$age_r[i] <- kc_data$age[i]
  } else {
    kc_data$age_r[i] <- 2015 - kc_data$yr_renovated[i]
    }
}
```

Examine the new variable "Age"

```
summary(kc_data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   18.00   40.00   43.99   64.00  115.00
```

```
par(mfrow = c(1,2))
hist(kc_data$age, col="dodgerblue", main="Property Age Distribution", xlab="Property Age")
plot(price ~ age, data=kc_data, cex=0.5, col="dodgerblue", main="Price-Age Relationship", xlab="Property Age", yl
ab="Property Price")
```

We can see that there are more modern than old houses in King County. Additionally, the largest price outliers are either very new or very old properties which indicates a quadratic price-age relationship (more on that later).

Furthermore, all variables regarding size of the property are measured in square foot, but we want to measure them in square meters. Therefore, we provide a function that converts square foot to square meter.

New function: Convert sqft to sqm

```
sqft_sqm <- function(sqft) {
  sqm <- round(sqft/10.76391)
  return(sqm)
}
```

Apply the function to all variables that are measured in square foot:

```
kc_data$sqm_living <- sqft_sqm(kc_data$sqft_living)
kc_data$sqm_lot <- sqft_sqm(kc_data$sqft_lot)
kc_data$sqm_above <- sqft_sqm(kc_data$sqft_above)
kc_data$sqm_basement <- sqft_sqm(kc_data$sqft_basement)
kc_data$sqm_living15 <- sqft_sqm(kc_data$sqft_living15)
kc_data$sqm_lot15 <- sqft_sqm(kc_data$sqft_lot15)
```

Examine the new variable “Square meters of living space”

```
summary(kc_data$sqm_living)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    27.0   133.0   177.0   193.2   237.0  1258.0
```

The dataset includes properties with a living space between 27 and 1258 square meters. The average living space is approximately 193 square meters.

```
par(mfrow = c(1,2))
hist(kc_data$sqm_living, col="red", main="Property Size Distribution", xlab="Square Meters")
plot(price ~ sqm_living, data=kc_data, cex=0.5, col="red", main="Price-Size Relationship", xlab="Property Size",
ylab="Property Price")
```

As expected there are more small than large properties in King County. The price-size relationship seems to be linear. More space leads on average to a higher house price.

Get an overview of all variables we now have:

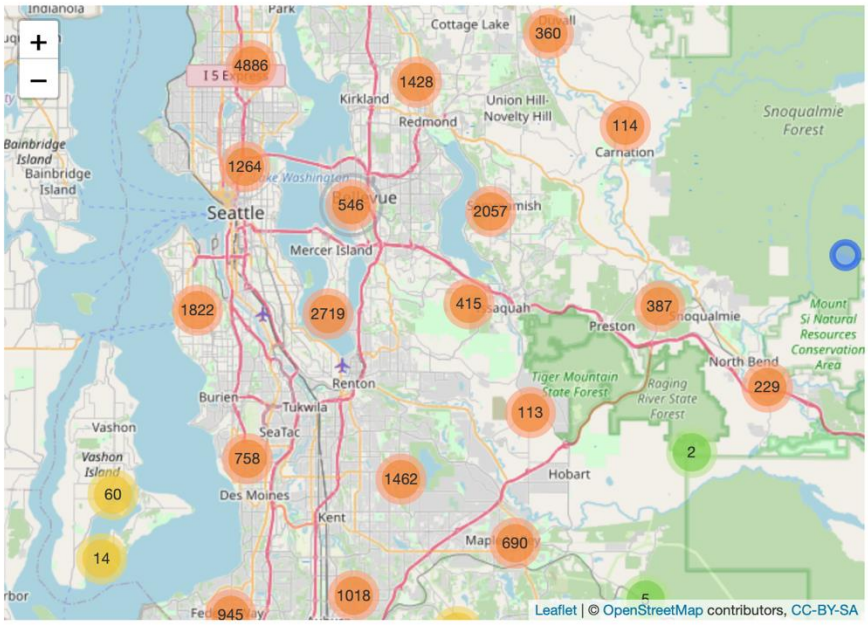
```
ls(kc_data)
```

```
##  [1] "age"           "age_r"         "bathrooms"     "bedrooms"
##  [5] "condition"     "floors"        "grade"         "lat"
##  [9] "long"          "price"         "sqft_above"    "sqft_basement"
## [13] "sqft_living"   "sqft_living15" "sqft_lot"      "sqft_lot15"
## [17] "sqm_above"     "sqm_basement"  "sqm_living"    "sqm_living15"
## [21] "sqm_lot"       "sqm_lot15"     "view"          "waterfront"
## [25] "yr_built"      "yr_renovated"  "zipcode"
```

Part 3: Visualize the location of the properties and their prices

Create an interactive map that shows the location of all properties:

```
m1 <- leaflet()
m1 <- setView(m1, lng = mean(kc_data$long), lat = mean(kc_data$lat), zoom = 10)
m1 <- addTiles(m1)
m1 <- addCircleMarkers(m1, lng = kc_data$long, lat = kc_data$lat,
                       clusterOptions = markerClusterOptions())
m1
```



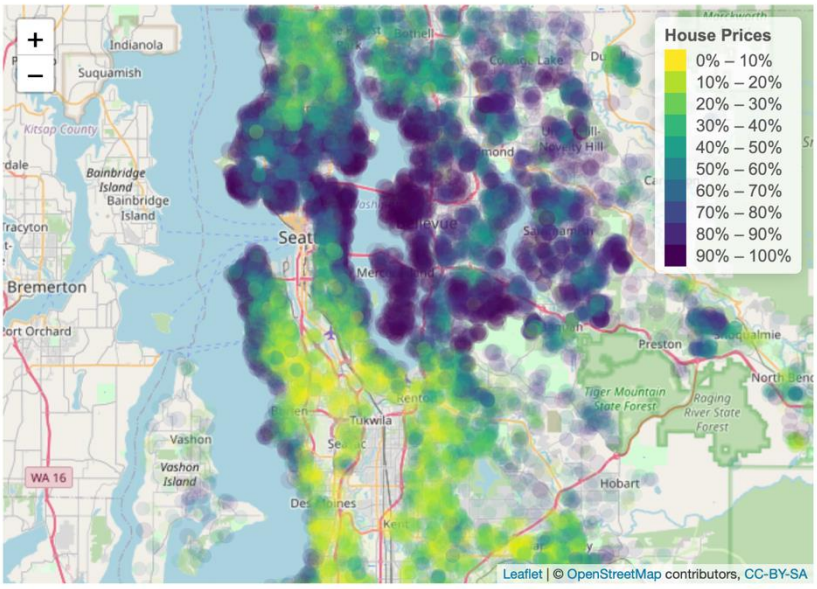
This map shows the density of the properties in King County.

Create an interactive map that shows the house price in every location (purple=expensive, yellow=cheap):

```
m2 <- leaflet()
m2 <- setView(m2, lng = mean(kc_data$long), lat = mean(kc_data$lat), zoom = 10)
m2 <- addTiles(m2)

pal.quantile <- colorQuantile("viridis",
                              domain = kc_data$price, reverse = TRUE, n = 10)
kc_data$price.colors.quant <- pal.quantile(kc_data$price)

m2 <- addCircleMarkers(m2, lng = kc_data$long,
                       lat = kc_data$lat, radius = log(kc_data$price/1000),
                       stroke = FALSE, fillOpacity = 0.1, fill = T,
                       fillColor =  kc_data$price.colors.quant)
m2 <- addLegend(m2, pal = pal.quantile, values = kc_data$price, opacity = 1, title = "House Prices")
m2
```



Yellow circles show the 10% of the properties with the lowest price, dark purple circles show the 10% of the properties with the highest prices. The map shows that the cheapest houses are located in the south of Seattle while the most expensive houses are located in the north and east of Seattle. Especially the regions Bellevue and Mercer Island show very high house prices.

Part 4: Estimation of various regression models which will be the basis for the house price calculator

Effect of the property age:

```
r1 <- lm(price ~ age, data=kc_data)
summary(r1)
```

```
##
## Call:
## lm(formula = price ~ age, data = kc_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -461709 -221337  -87006  104064 7201095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 569787.8     4490.9 126.875  < 2e-16 ***
## age           -675.1       84.9  -7.952 1.93e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 366600 on 21611 degrees of freedom
## Multiple R-squared:  0.002917,   Adjusted R-squared:  0.002871
## F-statistic: 63.23 on 1 and 21611 DF,  p-value: 1.93e-15
```

```
r2 <- lm(price ~ age + I(age^2), data=kc_data)
summary(r2)
```

```
##
## Call:
## lm(formula = price ~ age + I(age^2), data = kc_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -567832 -216629  -83529   99931 7074359
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 675334.905   6338.387  106.55   <2e-16 ***
## age          -7070.541    286.879  -24.65   <2e-16 ***
## I(age^2)        62.831      2.695   23.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 362100 on 21610 degrees of freedom
## Multiple R-squared:  0.02738,    Adjusted R-squared:  0.02729
## F-statistic: 304.1 on 2 and 21610 DF,  p-value: < 2.2e-16
```

The regression models for the variables “Age” and “Age^2” show that they have a significant impact (indicated by a very low p-value) on the house prices. This relationship seems to be quadratic because both terms age and age^2 are significant.

We now include all variables:

```
r3 <- lm(price ~ age + I(age^2) + age_r + I(age_r^2) + bedrooms + bathrooms + floors + waterfront + view + condit
ion + grade + sqm_living + sqm_lot + sqm_above + sqm_basement + sqm_living15 + sqm_lot15, data=kc_data)
summary(r3)
```

```
##
## Call:
## lm(formula = price ~ age + I(age^2) + age_r + I(age_r^2) + bedrooms +
##     bathrooms + floors + waterfront + view + condition + grade +
##     sqm_living + sqm_lot + sqm_above + sqm_basement + sqm_living15 +
##     sqm_lot15, data = kc_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1241332  -109113   -10278    89293  4420714
##
## Coefficients:
##                   Estimate    Std. Error t value Pr(>|t|)
## (Intercept)  -953698.09461   18199.04322 -52.404  < 2e-16 ***
## age             7065.54621     522.16539  13.531  < 2e-16 ***
## I(age^2)         -33.83650       4.83413  -6.999 2.64e-12 ***
## age_r          -5511.80071     514.33174 -10.716  < 2e-16 ***
## I(age_r^2)        52.46502       4.85330  10.810  < 2e-16 ***
## bedrooms       -37337.51951    2034.10444 -18.356  < 2e-16 ***
## bathrooms       41817.84536    3520.38893  11.879  < 2e-16 ***
## floors          12862.70290    4187.21898   3.072  0.00213 **
## waterfront     585524.50249   18551.84501  31.562  < 2e-16 ***
## view            44290.66523    2268.65605  19.523  < 2e-16 ***
## condition       23000.46811    2525.52164   9.107  < 2e-16 ***
## grade          119941.63700    2242.35570  53.489  < 2e-16 ***
## sqm_living       -101.55394    4543.38590  -0.022  0.98217
## sqm_lot             0.04181       0.55003   0.076  0.93941
## sqm_above        1828.77995    4543.44451   0.403  0.68731
## sqm_basement     1899.81288    4544.03417   0.418  0.67589
## sqm_living15      271.22297      38.62627   7.022 2.26e-12 ***
## sqm_lot15          -5.53440       0.84153  -6.577 4.92e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 215300 on 21595 degrees of freedom
## Multiple R-squared:  0.6563, Adjusted R-squared:  0.656
## F-statistic:  2426 on 17 and 21595 DF,  p-value: < 2.2e-16
```

Some of the variables regarding property size are not significant because there is a high correlation between them. We therefore exclude all of them apart from living space and overall property size.

```
r4 <- lm(price ~ age + I(age^2) + age_r + I(age_r^2) + bedrooms + bathrooms + floors + waterfront + view + condit
ion + grade + sqm_living + sqm_lot, data=kc_data)
summary(r4)
```

```
##
## Call:
## lm(formula = price ~ age + I(age^2) + age_r + I(age_r^2) + bedrooms +
##     bathrooms + floors + waterfront + view + condition + grade +
##     sqm_living + sqm_lot, data = kc_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1251282  -109367    -9928    89518  4370176
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -954450.132   18164.431 -52.545  < 2e-16 ***
## age            6931.852     522.536  13.266  < 2e-16 ***
## I(age^2)        -33.105       4.842  -6.838 8.26e-12 ***
## age_r         -5438.533     515.034 -10.560  < 2e-16 ***
## I(age_r^2)       52.294       4.861  10.757  < 2e-16 ***
## bedrooms     -36795.776    2035.366 -18.078  < 2e-16 ***
## bathrooms     41873.784    3471.253  12.063  < 2e-16 ***
## floors         9109.417    3890.116   2.342   0.0192 *
## waterfront   580243.487   18566.898  31.252  < 2e-16 ***
## view          46463.142    2224.151  20.890  < 2e-16 ***
## condition     22461.331    2524.378   8.898  < 2e-16 ***
## grade        124519.655    2131.542  58.418  < 2e-16 ***
## sqm_living     1835.735      35.133  52.251  < 2e-16 ***
## sqm_lot          -2.525       0.391  -6.459 1.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 215700 on 21599 degrees of freedom
## Multiple R-squared:  0.6549, Adjusted R-squared:  0.6547
## F-statistic:  3153 on 13 and 21599 DF,  p-value: < 2.2e-16
```

All of the included variables are now significant (at least at a 5% level). This model will be used as a basis for the house price calculator.

Part 5: House Price Calculator

Ask the user for all the required input parameters:

```
user_age <- as.numeric(readline(prompt="Please enter the age of your desired property in years: "))
```

```
## Please enter the age of your desired property in years:
```

```
user_age_r <- as.numeric(readline(prompt="Please enter the years since the last renovation: "))
```

```
## Please enter the years since the last renovation:
```

```
user_bedrooms <- as.numeric(readline(prompt="Please enter the number of bedrooms: "))
```

```
## Please enter the number of bedrooms:
```

```
user_bathrooms <- as.numeric(readline(prompt="Please enter the number of bathrooms: "))
```

```
## Please enter the number of bathrooms:
```

```
user_floors <- as.numeric(readline(prompt="Please enter the number of floors: "))
```

```
## Please enter the number of floors:
```

```
user_waterfront <- as.numeric(readline(prompt="Please indicate if your desired property is located at the waterfr
ont (1 for yes, 0 for no): "))
```

```
## Please indicate if your desired property is located at the waterfront (1 for yes, 0 for no):
```

```
user_view <- as.numeric(readline(prompt="Please indicate on a scale from 0 to 4 how good the view from your prope
rty is (0 = no view, 4 = perfect view): "))
```

```
## Please indicate on a scale from 0 to 4 how good the view from your property is (0 = no view, 4 = perfect view)
:
```

```
user_condition <- as.numeric(readline(prompt="Please indicate the condition of the house on a scale from 1 to 5 (
1 = bad, 5 = perfect): "))
```

```
## Please indicate the condition of the house on a scale from 1 to 5 (1 = bad, 5 = perfect):
```

```
user_grade <- as.numeric(readline(prompt="Please indicate on a scale from 1 to 13 how good the building construct
ion and design are (1 = bad, 7 = average, 13 = perfect): "))
```

```
## Please indicate on a scale from 1 to 13 how good the building construction and design are (1 = bad, 7 = averag
e, 13 = perfect):
```

```
user_sqm_living <- as.numeric(readline(prompt="Please enter the size of the interior living space in square meter
s: "))
```

```
## Please enter the size of the interior living space in square meters:
```

```
user_sqm_lot <- as.numeric(readline(prompt="Please enter the size of the overall land space in square meters: "))
```

```
## Please enter the size of the overall land space in square meters:
```

Store user's input in a new data frame:

```
userhouse <- data.frame(age=user_age, age_r=user_age_r, bedrooms=user_bedrooms, bathrooms=user_bathrooms, floors=
user_floors, waterfront=user_waterfront, view=user_view, condition=user_condition, grade=user_grade, sqm_living=u
ser_sqm_living, sqm_lot=user_sqm_lot)
```

Predict the house price based on the model:

```
predict(r4, newdat = userhouse, interval = "prediction", level = 0.25)
```

```
##   fit lwr upr
## 1  NA  NA  NA
```

The output of the calculator contains three numbers. "Fit" is the estimated house price and "Lwr" and "Upr" define an error margin for the output. They provide an interval in which the house price will probably lie.

Now we test our calculator and compute the property value for our sample buyer who would like to estimate the price he has to pay in order to acquire his dream home. The house characteristics are listed below.

Example property: 50 years old, 10 years since last renovation, 3 bedrooms, 2 bathrooms, 2 floors, no waterfront, average view (2), condition is good (4), construction grade 7, 120 square meters of living space, 300 square meters of land space

```
examplehouse <- data.frame(age=50, age_r=10, bedrooms=3, bathrooms=2, floors=2, waterfront=0, view=2, condition=4
, grade=7, sqm_living=120, sqm_lot=300)
predict(r4, newdat = examplehouse, interval = "prediction", level = 0.25)
```

```
##        fit      lwr    upr
## 1 525742.9 456912.8 594573
```

The dream home of our sample buyer has an estimated house price of approximately 525000 USD.

The house price calculator can be used for any buyer or seller who wants to estimate the property value of an object he or she wants to buy or sell given certain house features.

6. Sources

Dataset: House Prices from King County, USA:

https://www.kaggle.com/harlfoxem/housesalesprediction