

Maestría en Inteligencia Analítica de Datos.**Estudiantes:** Cristian Camilo Ospina Alzate

Daniel Borda

Federico Higuera

Carlos Adrián Alarcón

Materia: Modelos de Análisis Estadístico**Trayectoria I – Ciclo II****ENTREGA 1 PROYECTO FINAL – ANÁLISIS EXPLORATORIO DE DATOS****CONTEXTO DEL PROBLEMA:**

El gerente de la empresa inmobiliaria Casa Roble está interesado en predecir el precio de venta de una vivienda a partir de un conjunto de variables que tiene a su disposición, tales como año de construcción, tamaño, número de parqueaderos, entre otras.

Para el efecto cuenta con alguna información que ha sido recolectada durante los últimos diez años sobre el precio de venta de apartamentos y algunas variables explicativas.

En la evaluación del modelo de regresión planteado, se tomarán en cuenta los siguientes resultados propios de su desarrollo:

- *Elección de las variables independientes del modelo*
- *Estructura del modelo*
- *Estimación de los parámetros del modelo*
- *Verificación de los supuestos del modelo de regresión*
- *Bondad de ajuste del modelo propuesto*
- *Cálculo de intervalos de confianza para los parámetros del modelo*
- *Prueba de hipótesis sobre los parámetros del modelo*
- *Interpretación de los coeficientes del modelo*

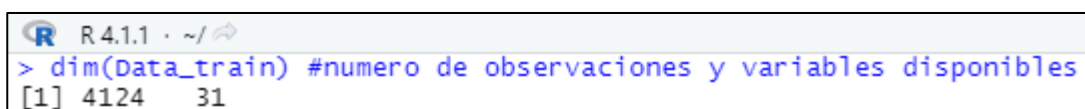
Problema de clasificación: Marketing bancario

.....

Problema de regresión: Valor de la vivienda

A continuación, en el presente informe se muestra una exploración de los datos del archivo “Train Real State” el cual corresponde al problema de regresión del valor de la vivienda, perteneciente al proyecto final de la materia.

En la siguiente imagen podemos apreciar la cantidad de las variables y la cantidad de observaciones por cada una de ellas:



```
R 4.1.1 ~ / > dim(Data_train) #numero de observaciones y variables disponibles [1] 4124 31
```

Donde tenemos un total de 31 variables y 4124 observaciones, ahora es de nuestro interés conocer el nombre de cada una de estas variables:

[1] "X"	"SalePrice"	"YearBuilt"
[4] "YrSold"	"MonthSold"	"Size.sqf."
[7] "Floor"	"HallwayType"	"HeatingType"
[10] "AptManagetType"	"N_Parkinglot.Ground."	"N_Parkinglot.Basement."
[13] "TimeToBusStop"	"TimeToSubway"	"N_APT"
[16] "N_manager"	"N_elevators"	"SubwayStation"
[19] "N_FacilitiesNearBy.PublicOffice."	"N_FacilitiesNearBy.Hospital."	"N_FacilitiesNearBy.Dpartmentstore."
[22] "N_FacilitiesNearBy.Mall."	"N_FacilitiesNearBy.ETC."	"N_FacilitiesNearBy.Park."
[25] "N_SchoolNearBy.Elementary."	"N_SchoolNearBy.Middle."	"N_SchoolNearBy.High."
[28] "N_SchoolNearBy.University."	"N_FacilitiesInApt"	"N_FacilitiesNearBy.Total."
[31] "N_SchoolNearBy.Total."		

En la figura anterior podemos apreciar el nombre de cada una de las variables de la base de datos, ahora es de vital interés saber cuáles de estas variables son categóricas y cuales son numéricas, lo anterior lo podemos apreciar en la siguiente figura:

```
R 4.1.1 ~ /
> str(Data_train)
'data.frame': 4124 obs. of 31 variables:
 $ X                : int  1 2 3 5 7 8 10 11 12 13 ...
 $ SalePrice        : int  141592 51327 48672 221238 78318 61946 83185 168141 153982 200884 ...
 $ YearBuilt        : int  2006 1985 1985 1993 1992 1993 1992 1986 1986 2007 ...
 $ YrSold           : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
 $ MonthSold        : int  8 8 8 8 8 8 8 8 8 8 ...
 $ Size.sqf         : int  814 587 587 1761 644 644 644 1377 914 868 ...
 $ Floor            : int  3 8 6 3 2 10 13 4 11 18 ...
 $ HallwayType      : chr  "terraced" "corridor" "corridor" "mixed" ...
 $ HeatingType      : chr  "individual_heating" "individual_heating" "individual_heating" "individual_heating" ...
 $ AptManagetType   : chr  "management_in_trust" "self_management" "self_management" "management_in_trust" ...
 $ N_Parkinglot.Ground. : int  111 80 80 523 142 523 142 713 713 0 ...
 $ N_Parkinglot.Basement. : int  184 76 76 536 79 536 79 0 0 1270 ...
 $ TimeToBusStop    : chr  "5min~10min" "0~5min" "0~5min" "0~5min" ...
 $ TimeToSubway     : chr  "10min~15min" "5min~10min" "5min~10min" "15min~20min" ...
 $ N_APT            : int  3 1 1 8 3 8 3 7 7 7 ...
 $ N_manager        : int  3 2 2 8 4 8 4 8 8 14 ...
 $ N_elevators       : int  0 2 2 20 8 20 8 27 27 16 ...
 $ SubwayStation    : chr  "kyungbuk_uni_hospital" "Daegu" "Daegu" "Myung-duk" ...
 $ N_FacilitiesNearBy.PublicOffice. : int  2 5 5 6 5 6 5 5 5 3 ...
 $ N_FacilitiesNearBy.Hospital. : int  1 1 1 2 1 2 1 1 1 1 ...
 $ N_FacilitiesNearBy.Dpartmentstore. : int  1 2 2 0 1 0 1 1 1 2 ...
 $ N_FacilitiesNearBy.Mall. : int  1 1 1 1 1 1 1 0 0 1 ...
 $ N_FacilitiesNearBy.ETC. : int  1 2 2 5 1 5 1 1 1 0 ...
 $ N_FacilitiesNearBy.Park. : int  0 1 1 0 0 0 0 1 1 2 ...
 $ N_SchoolNearBy.Elementary. : int  3 2 2 4 3 4 3 3 3 3 ...
 $ N_SchoolNearBy.Middle. : int  2 1 1 3 3 3 3 1 1 3 ...
 $ N_SchoolNearBy.High. : int  2 1 1 5 4 5 4 1 1 2 ...
 $ N_SchoolNearBy.University. : int  2 0 0 5 4 5 4 1 1 2 ...
 $ N_FacilitiesInApt : int  5 3 3 4 3 4 3 4 4 10 ...
 $ N_FacilitiesNearBy.Total. : int  6 12 12 14 9 14 9 9 9 9 ...
 $ N_SchoolNearBy.Total. : int  9 4 4 17 14 17 14 6 6 10 ...
```

En este orden de ideas, en la figura anterior podemos apreciar que tenemos un total de 6 variables categóricas y el resto son numéricas. Teniendo en cuenta lo anterior y para realizar la exploración de los datos, primero que todo se presentará el análisis para las variables numéricas y en una segunda sección se presentarán el de las variables categóricas.

I. ANÁLISIS EXPLORATORIO PARA LAS VARIABLES NUMÉRICAS

Primero que todo para entrar en contexto y obtener conocimiento acerca de los atributos de las variables numéricas que contiene la base de datos, se mostrará en la siguiente figura algunos datos de interés para cada variable tales como:

- Valor mínimo
- Cuartil del 25%
- Mediana
- Promedio
- Cuartil del 75%
- Valor máximo

```
R 4.1.1 ~ /
```

```
> summary(numericas)
```

SalePrice	YearBuilt	YrSold	MonthSold	Size.sqf.	Floor	N_Parkinglot.Ground.
Min. : 32743	Min. :1978	Min. :2007	Min. : 1.000	Min. : 135.0	Min. : 1.00	Min. : 0.0
1st Qu.:145464	1st Qu.:1993	1st Qu.:2010	1st Qu.: 3.000	1st Qu.: 644.0	1st Qu.: 6.00	1st Qu.: 11.0
Median :207964	Median :2006	Median :2013	Median : 6.000	Median : 910.0	Median :11.00	Median :100.0
Mean :221688	Mean :2003	Mean :2013	Mean : 6.163	Mean : 955.9	Mean :12.06	Mean :194.7
3rd Qu.:291150	3rd Qu.:2008	3rd Qu.:2015	3rd Qu.: 9.000	3rd Qu.:1149.0	3rd Qu.:17.00	3rd Qu.:249.0
Max. :585840	Max. :2015	Max. :2017	Max. :12.000	Max. :2337.0	Max. :43.00	Max. :713.0

N_Parkinglot.Basement.	N_APT	N_manager	N_elevators	N_FacilitiesNearBy.PublicOffice.
Min. : 0.0	Min. : 1.00	Min. : 1.000	Min. : 0.00	Min. :0.000
1st Qu.: 184.0	1st Qu.: 3.00	1st Qu.: 5.000	1st Qu.: 5.00	1st Qu.:3.000
Median : 536.0	Median : 7.00	Median : 6.000	Median :11.00	Median :5.000
Mean : 571.6	Mean : 5.63	Mean : 6.305	Mean :11.12	Mean :4.139
3rd Qu.: 798.0	3rd Qu.: 8.00	3rd Qu.: 8.000	3rd Qu.:16.00	3rd Qu.:5.000
Max. :1321.0	Max. :13.00	Max. :14.000	Max. :27.00	Max. :7.000

N_FacilitiesNearBy.Hospital.	N_FacilitiesNearBy.Dpartmentstore.	N_FacilitiesNearBy.Mall.	N_FacilitiesNearBy.ETC.
Min. :0.000	Min. :0.0000	Min. :0.0000	Min. :0.000
1st Qu.:1.000	1st Qu.:0.0000	1st Qu.:1.0000	1st Qu.:0.000
Median :1.000	Median :1.0000	Median :1.0000	Median :1.000
Mean :1.295	Mean :0.8887	Mean :0.9391	Mean :1.937
3rd Qu.:2.000	3rd Qu.:2.0000	3rd Qu.:1.0000	3rd Qu.:5.000
Max. :2.000	Max. :2.0000	Max. :2.0000	Max. :5.000

N_FacilitiesNearBy.Park.	N_SchoolNearBy.Elementary.	N_SchoolNearBy.Middle.	N_SchoolNearBy.High.	N_SchoolNearBy.University.
Min. :0.0000	Min. :0.000	Min. :0.000	Min. :0.000	Min. :0.000
1st Qu.:0.0000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:1.000	1st Qu.:2.000
Median :1.0000	Median :3.000	Median :3.000	Median :2.000	Median :2.000
Mean :0.6479	Mean :3.025	Mean :2.425	Mean :2.661	Mean :2.764
3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:3.000	3rd Qu.:4.000	3rd Qu.:4.000
Max. :2.0000	Max. :6.000	Max. :4.000	Max. :5.000	Max. :5.000

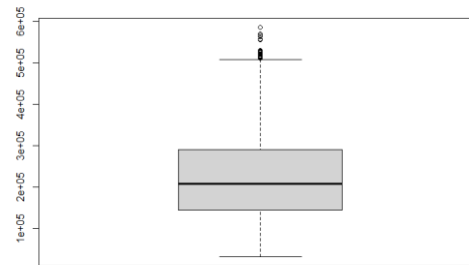
N_FacilitiesInApt	N_FacilitiesNearBy.Total.	N_SchoolNearBy.Total.
Min. : 1.000	Min. : 0.000	Min. : 0.00
1st Qu.: 4.000	1st Qu.: 8.000	1st Qu.: 7.00
Median : 5.000	Median : 9.000	Median :10.00
Mean : 5.828	Mean : 9.846	Mean :10.87
3rd Qu.: 7.000	3rd Qu.:13.000	3rd Qu.:15.00
Max. :10.000	Max. :16.000	Max. :17.00

```
> |
```

Se debe tener presente que la cantidad de variables que se presentaron en la figura anterior es de 24 ya que las categóricas no son de nuestro interés en este momento.

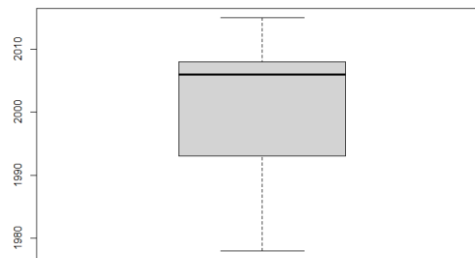
Ahora el interés en la exploración de los datos esta en identificar las observaciones cuyo valor difiere considerablemente respecto al conjunto de datos y pueden distorsionar el resultado del análisis. Para el análisis univariado utilizaremos el rango intercuantil para detectar datos atípicos, este método considera un dato como atípico si se encuentra 1.5 veces por encima o por debajo del rango de los datos, por ende, en las siguientes figuras se presentará lo mencionado anteriormente para cada una de las variables numéricas:

1. SalePrice



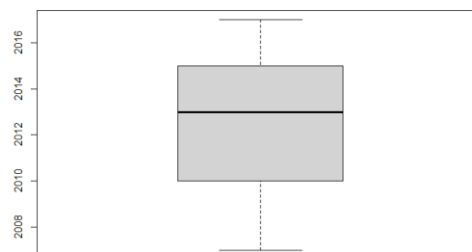
```
R 4.1.1 ~/>
> boxplot(numericas$SalePrice)$out
[1] 556637 527433 513274 570796 530973 566371 530973 526548 517699 517699 530973 529203 526548 522123 564601 529203 566371
[18] 515929 585840 515044 526548 522123 557522 570796 511504
> |
```

2. YearBuilt



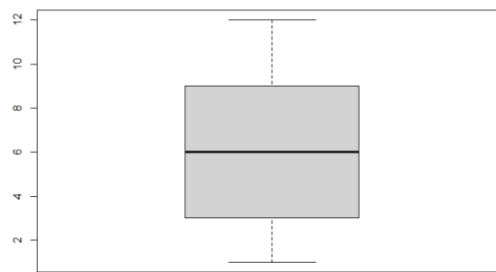
```
> boxplot(numericas$YearBuilt)$out
numeric(0)
> |
```

3. YrSold



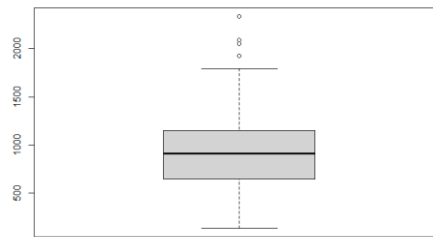
```
numeric(0)
> boxplot(numericas$Yrsold)$out
numeric(0)
> |
```

4. MonthSold



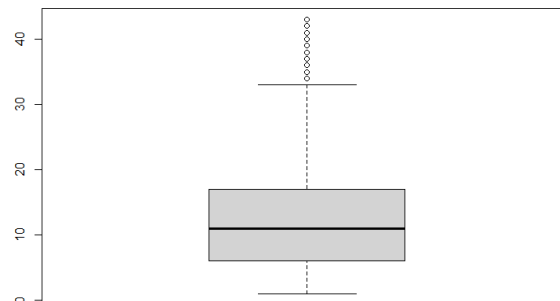
```
> boxplot(numericas$MonthSold)$out
numeric(0)
> |
```

5. Size.sqf.



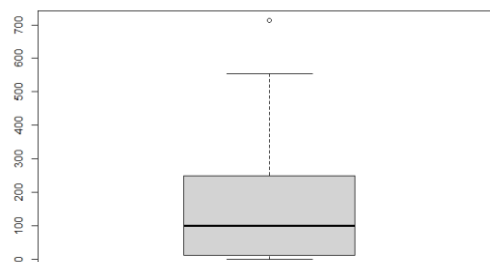
```
> boxplot(numericas$Size.sqf.)$out
[1] 2337 2337 2337 2056 2056 1928 1928 2337 1928 2337 1928 2337 2056 2056 2337 2337 2337 2056 2337 2337 2056 2337 2056 2092 2092
[26] 2056 2056 2337 2337 1928 2337 2056 2337 1928 2337 2056 1928 2056 2337 1928 1928 1928 1928 1928 1928 1928 2337 2337
[51] 2337 1928 1928 1928 2337 1928 2056 2056 2337 1928 2056 2056 1928 2056 2337 2056 2056 1928 2092 2337 1928 1928 1928 2056 1928
[76] 1928 2056 2056 1928 2337 1928 2056 1928 1928 2092 1928 1928 1928
> |
```

6. Floor



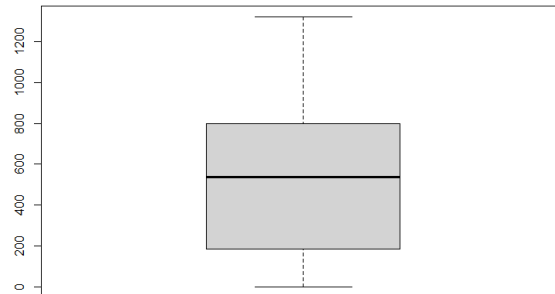
```
> boxplot(numericas$Floor)$out
[1] 39 43 38 35 43 35 42 42 43 37 38 39 40 36 37 42 37 41 34 35 39 40 42 43 41 41 42 36 40 35 37 35 38 35 35 41 35 41 37 38 41
[42] 42
> |
```

7. N_Parkinglot.Ground.



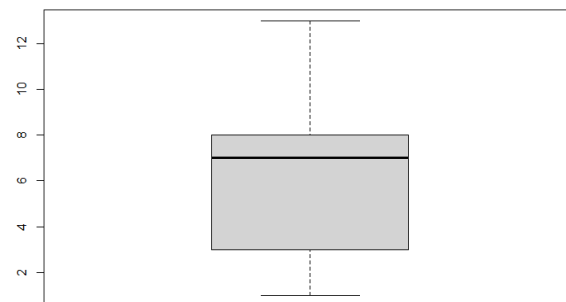
```
> boxplot(numericas$N_Parkinglot.Ground.)$out
[1] 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713
[32] 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713
[63] 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713
[94] 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713
[125] 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713
[156] 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713
[187] 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713 713
> |
```

8. N_Parkinglot.Basement.



```
> boxplot(numericas$N_Parkinglot.Basement.)$out
numeric(0)
> |
```

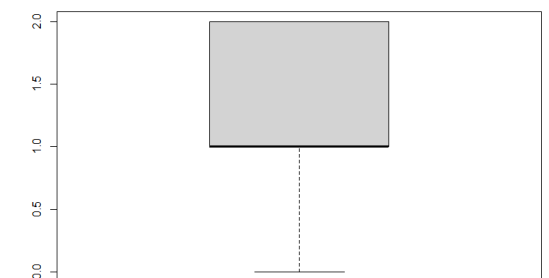
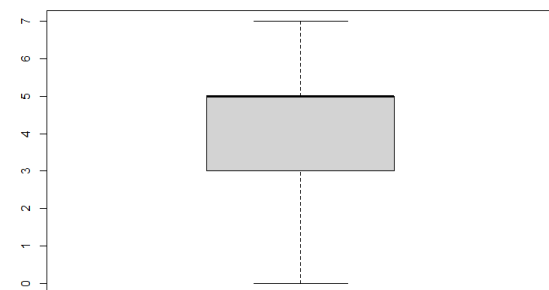
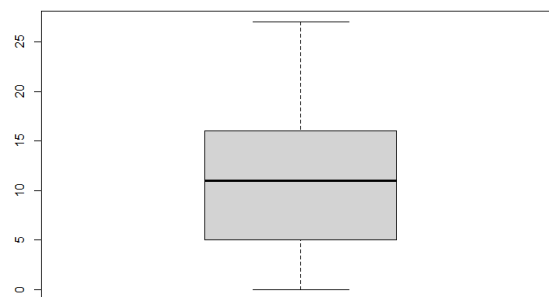
9. N_APT



```
> boxplot(numericas$N_APT)$out
numeric(0)
> |
```

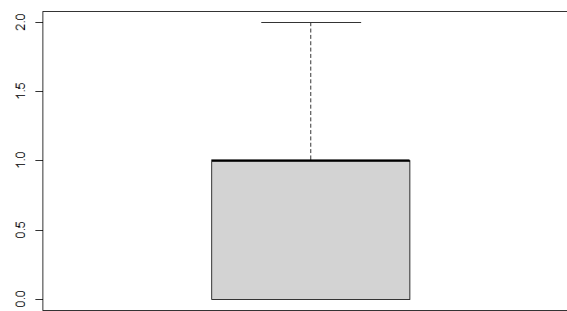
10. N_manager





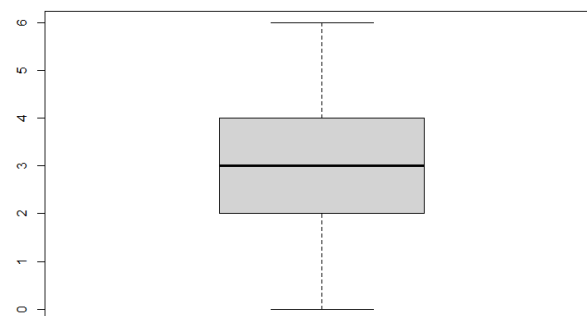
```
> boxplot(numericas$N_FacilitiesNearBy.Hospital.)$out
numeric(0)
> |
```

17. N_FacilitiesNearBy.Park.



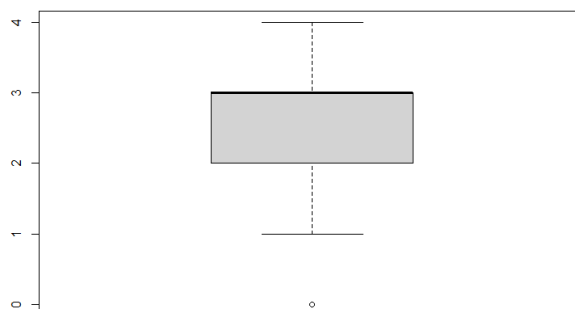
```
> boxplot(numericas$N_FacilitiesNearBy.Park.)$out
numeric(0)
> |
```

18. N_SchoolNearBy.Elementary.

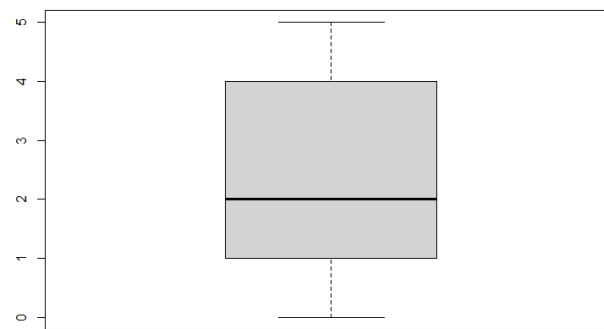


```
> boxplot(numericas$N_SchoolNearBy.Elementary.)$out
numeric(0)
>
```

19. N_SchoolNearBy.Middle.

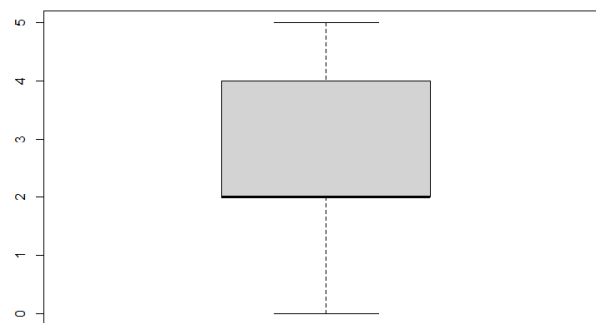
[illegible]

20. N_SchoolNearBy.High.



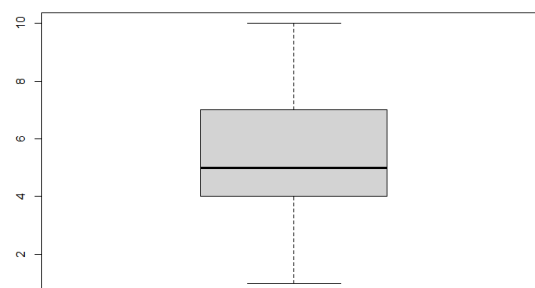
```
> boxplot(numericas$N_SchoolNearBy.High.)$out
numeric(0)
> |
```

21. N_SchoolNearBy.University.



```
> boxplot(numericas$N_SchoolNearBy.University.)$out
numeric(0)
> |
```

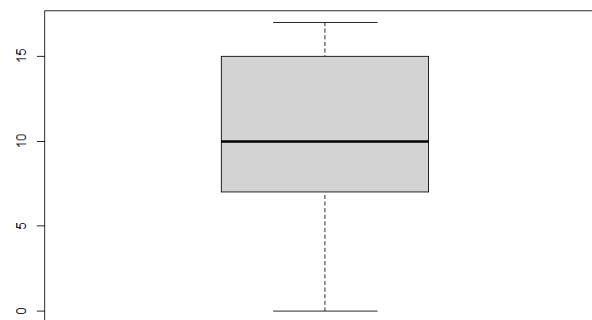
22. N_FacilitiesInApt



```
> boxplot(numericas$N_FacilitiesInApt)$out
numeric(0)
> |
```

23. N_FacilitiesNearBy.Total.

24. N_SchoolNearBy.Total.



Por último, una vez identificadas las medidas de tendencia central, es demasiado importante conocer además las medidas de dispersión tales como la varianza, la desviación estándar y el coeficiente de variación, para esto; en las siguientes tablas se presentará cada una de estas medidas para cada variable numéricas:

	N_FacilitiesNearBy.PublicOffice.	N_FacilitiesNearBy.Hospital.	N_FacilitiesNearBy.Dpartment.
VARIANZA	3,19	0,23	0,65
DESV. EST.	1,79	0,48	0,81

COEF. VAR.	0,431555055	0,371780661	0,906851712
------------	-------------	-------------	-------------

	N_FacilitiesNearBy.Mall.	N_FacilitiesNearBy.ETC.	N_FacilitiesNearBy.Park.
VARIANZA	0,16	4,85	0,43
DESV. EST.	0,40	2,20	0,66
COEF. VAR.	0,423346648	1,136604958	1,016606864

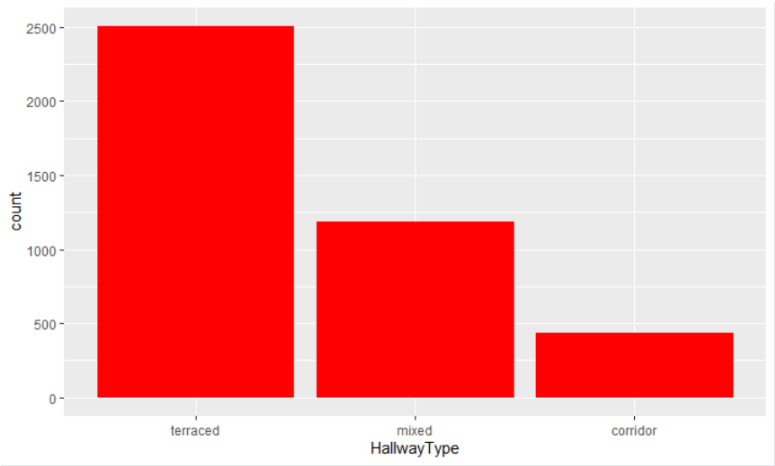
	N_SchoolNearBy.Elementary.	N_SchoolNearBy.Middle.	N_SchoolNearBy.High.
VARIANZA	0,91	1,07	2,40
DESV. EST.	0,95	1,04	1,55
COEF. VAR.	0,314939984	0,427199466	0,582087563

	N_SchoolNearBy.University.	N_FacilitiesInApt	N_FacilitiesNearBy.Total.	N_SchoolNearBy.Total.
VARIANZA	2,21	5,38	11,93	19,59
DESV. EST.	1,49	2,32	3,45	4,43
COEF. VAR.	0,537652335	0,398135315	0,350752464	0,406947759

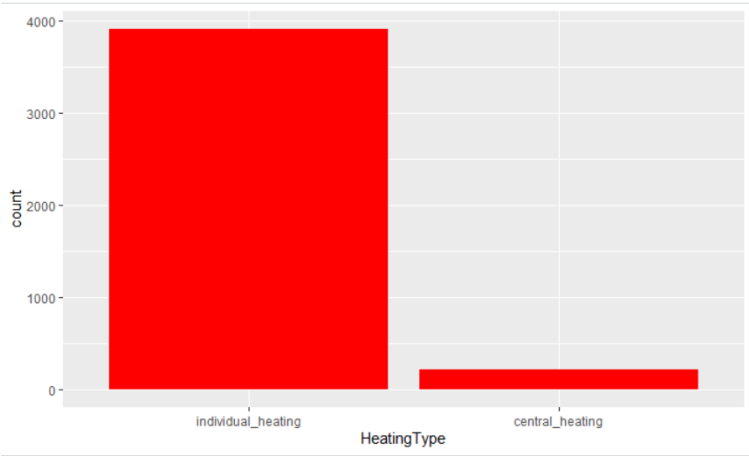
Variables Catagoricas.

En la siguiente parte se muestra un resumen de frecuencias de las variables de naturaleza categórica que se encuentran en la base:

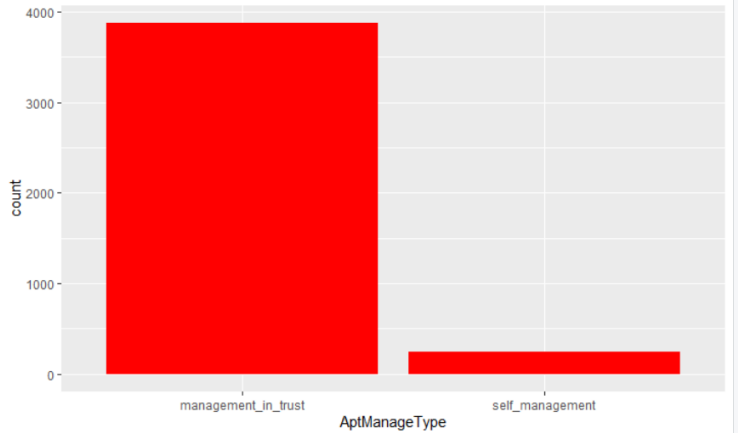
Hallway Tipe:



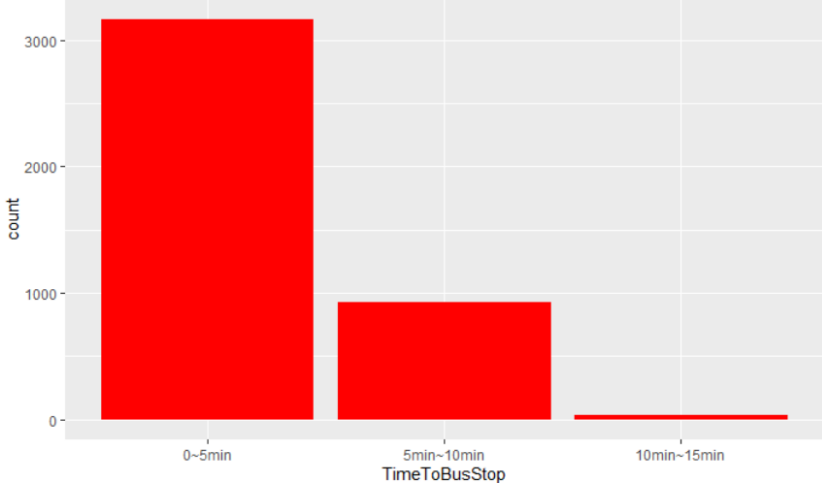
Heating Type:



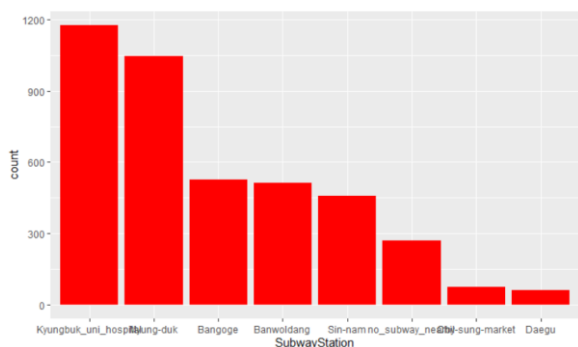
AptManage Type:



Time to Bus Stop:



Subway Station:



.....

Problema de clasificación : Marketing Bancario

A continuación, se explora la relación entre las variables de tipo categórico haciendo uso de **tablas de contingencia**, en donde se buscan indicios de diferencias entre las categorías de una variable respecto al desempeño de una segunda, a través de una examinación a lo largo de cada fila el porcentaje de las observaciones que se clasifican en cada una de sus celdas frente al total de su respectiva columna, Una diferencia marcada entre esos porcentajes sugiere la presencia de una relación entre las dos variables.

Para dichas comparaciones se establecieron dos grupos de categorías, la primera de ellas, la que a juicio del equipo se consideran factores que podrían o no influir sobre la segunda de estas, que se refieren a la adquisición de productos bancarios y otras variables de interés como el balance en euros.

Variables independientes	Variables independientes
<ul style="list-style-type: none">• Nivel de educación• Estatus civil ('Marital.Status')• Trabajo ('Job')• Rango de edad ('Rango de edad'*) <p>*Generada por el equipo</p>	<ul style="list-style-type: none">• ¿Cuenta con crédito? ('Credit')• ¿Tiene préstamo de vivienda? ('Housing.Loan')• ¿Tiene préstamo personal? ('Personal.Loan')• Balance ('Balance..euros.')

Se realiza el cruce y análisis de la tabla de contingencia para cada par de variables independientes y dependientes.

Educacion	Crédito			
	Crédito			Total general
	Educación	yes	no	
	primary	1.59%	98.41%	100.00%
	secondary	1.88%	98.12%	100.00%
Marital_Status	Credit			
	Marital_Status	yes	no	Total general
	divorced	2.58%	97.42%	100.00%
	married	1.55%	98.45%	100.00%
	single	1.69%	98.31%	100.00%
Job	Credit			
	Job	yes	no	Total general
	admin.	1.71%	98.29%	100.00%
	blue-collar	1.65%	98.35%	100.00%
	entrepreneur	2.62%	97.38%	100.00%
Edad	Credit			
	Edad	yes	no	Total general
	-30	1.88%	98.12%	100.00%
	30 - 39	1.46%	98.54%	100.00%
	40 - 49	2.33%	97.67%	100.00%
Educacion	Housing_Loan			
	Educación	yes	no	Total general
	primary	56.88%	43.12%	100.00%
	secondary	61.19%	38.81%	100.00%
	tertiary	49.40%	50.60%	100.00%
Marital_Status	Housing_Loan			
	Marital_Status	yes	no	Total general
	divorced	57.43%	42.57%	100.00%
	married	56.70%	43.30%	100.00%
	single	55.52%	44.48%	100.00%
Job	Housing_Loan			
	Job	yes	no	Total general
	admin.	0.621912603	0.378087397	1
	blue-collar	0.733917144	0.266082856	1
	entrepreneur	0.583333333	0.416666667	1
Edad	Housing_Loan			
	Edad	yes	no	Total general
	-30	60.28%	39.72%	100.00%
	30 - 39	63.69%	36.31%	100.00%
	40 - 49	59.16%	40.84%	100.00%
Educacion	Personal_Loan			
	Educación	yes	no	Total general
	primary	15.45%	84.55%	100.00%
	secondary	18.48%	81.52%	100.00%
	tertiary	12.67%	87.33%	100.00%
Marital_Status	Personal_Loan			
	Marital_Status	yes	no	Total general
	divorced	17.88%	82.12%	100.00%
	married	17.11%	82.89%	100.00%
	single	12.41%	87.59%	100.00%
Job	Personal_Loan			
	Job	yes	no	Total general
	admin.	19.19%	80.81%	100.00%
	blue-collar	17.11%	82.89%	100.00%
	entrepreneur	20.71%	79.29%	100.00%
Edad	Personal_Loan			
	Edad	yes	no	Total general
	-30	15.48%	84.52%	100.00%
	30 - 39	15.63%	84.37%	100.00%
	40 - 49	17.50%	82.50%	100.00%

Se detallarán a continuación, los hallazgos más relevantes del análisis.

Housing_Loan			
Educación			Total general
primary	56.88%	43.12%	100.00%
secondary	61.19%	38.81%	100.00%
tertiary	49.40%	50.60%	100.00%
unknown	44.10%	55.90%	100.00%
Total general	56.46%	43.54%	100.00%

Personal_Loan			
Educación			Total general
primary	15.45%	84.55%	100.00%
secondary	18.48%	81.52%	100.00%
tertiary	12.67%	87.33%	100.00%
unknown	7.08%	92.92%	100.00%
Total general	15.89%	84.11%	100.00%

Existe una aparente relación entre el nivel de estudio de la población y su intención por adquirir un préstamo para vivienda o personal. Se detecta una proporción significativamente mayor para la población con estudios secundarios comparada con los otros grupos.

Credit			
Marital_Status			Total general
divorced	2.58%	97.42%	100.00%
married	1.55%	98.45%	100.00%
single	1.69%	98.31%	100.00%
Total general	1.71%	98.29%	100.00%

Se detecta una mayor intención por las personas cuyo estado civil es ‘divorciado’ en obtener un crédito, comparado con las personas catalogadas en otros grupos.

% Personal_Loan			
Marital_Status	yes	no	Total general
divorced	17.88%	82.12%	100.00%
married	17.11%	82.89%	100.00%
single	12.41%	87.59%	100.00%
Total general	15.89%	84.11%	100.00%

Se sospecha de una menor intención por parte de las personas ‘solteras’ en obtener un préstamo personal.

% Credit			
Job	yes	no	Total general
admin.	1.71%	98.29%	100.00%
blue-collar	1.65%	98.35%	100.00%
entrepreneur	2.62%	97.38%	100.00%
housemaid	2.69%	97.31%	100.00%
management	1.78%	98.22%	100.00%
retired	1.06%	98.94%	100.00%
self-employed	1.76%	98.24%	100.00%
services	1.83%	98.17%	100.00%
student	0.00%	100.00%	100.00%
technician	1.68%	98.32%	100.00%
unemployed	1.89%	98.11%	100.00%
unknown	1.18%	98.82%	100.00%
Total general	1.71%	98.29%	100.00%

% Housing_Loan			
Job	yes	no	Total general
admin.	0.621912603	0.378087397	1
blue-collar	0.733917144	0.266082856	1
entrepreneur	0.583333333	0.416666667	1
housemaid	0.298387097	0.701612903	1
management	0.50174581	0.49825419	1
retired	0.193939394	0.806060606	1
self-employed	0.469162996	0.530837004	1
services	0.675810474	0.324189526	1
student	0.322097378	0.677902622	1
technician	0.552393273	0.447606727	1
unemployed	0.469002695	0.530997305	1
unknown	0.094117647	0.905882353	1
Total general	56.46%	43.54%	100.00%

Existe una aparente relación entre el trabajo que desempeña un individuo de la población y su interés por obtener cualquiera de los productos ofrecidos por la entidad bancaria. Adicionalmente, dicha variable posiblemente influencia sobre el promedio del balance de la población.

% Personal_Loan			
Job	yes	no	Total general
admin.	19.19%	80.81%	100.00%
blue-collar	17.11%	82.89%	100.00%
entrepreneur	20.71%	79.29%	100.00%
housemaid	12.37%	87.63%	100.00%
management	12.29%	87.71%	100.00%
retired	13.33%	86.67%	100.00%
self-employed	14.32%	85.68%	100.00%
services	21.11%	78.89%	100.00%
student	1.87%	98.13%	100.00%
technician	17.85%	82.15%	100.00%
unemployed	8.36%	91.64%	100.00%
unknown	2.35%	97.65%	100.00%
Total general	15.89%	84.11%	100.00%

Educación		romedio de Balance..euros
retired		2096.778788
management		1700.127793
self-employed		1636.614537
unemployed		1625.345013
entrepreneur		1492.416667
housemaid		1368.569892
unknown		1356.952941
technician		1149.527814
student		1143.992509
admin.		1129.748575
blue-collar		1060.708656
services		1034.921862
Total general		1327.114871

Se percibe una diferencia en la obtención de créditos de acuerdo con el rango de edad de la población, siendo aquellos de '40 a 49' años, quienes muestran mayor interés por este tipo de productos.

% Credit			
Edad	yes	no	Total general
-30	1.88%	98.12%	100.00%
30 - 39	1.46%	98.54%	100.00%
40 - 49	2.33%	97.67%	100.00%
50 - 59	1.48%	98.52%	100.00%
60 - 69	0.79%	99.21%	100.00%
70 - 79	0.78%	99.22%	100.00%
80 +	0.00%	100.00%	100.00%
Total general	1.71%	98.29%	100.00%

% Housing_Loan			
Edad	yes	no	Total general
-30	60.28%	39.72%	100.00%
30 - 39	63.69%	36.31%	100.00%
40 - 49	59.16%	40.84%	100.00%
50 - 59	44.00%	56.00%	100.00%
60 - 69	18.73%	81.27%	100.00%
70 - 79	3.10%	96.90%	100.00%
80 +	0.00%	100.00%	100.00%
Total general	56.46%	43.54%	100.00%

% Personal_Loan			
Edad	yes	no	Total general
-30	15.48%	84.52%	100.00%
30 - 39	15.63%	84.37%	100.00%
40 - 49	17.50%	82.50%	100.00%
50 - 59	16.61%	83.39%	100.00%
60 - 69	8.18%	91.82%	100.00%
70 - 79	1.55%	98.45%	100.00%
80 +	0.00%	100.00%	100.00%
Total general	15.89%	84.11%	100.00%

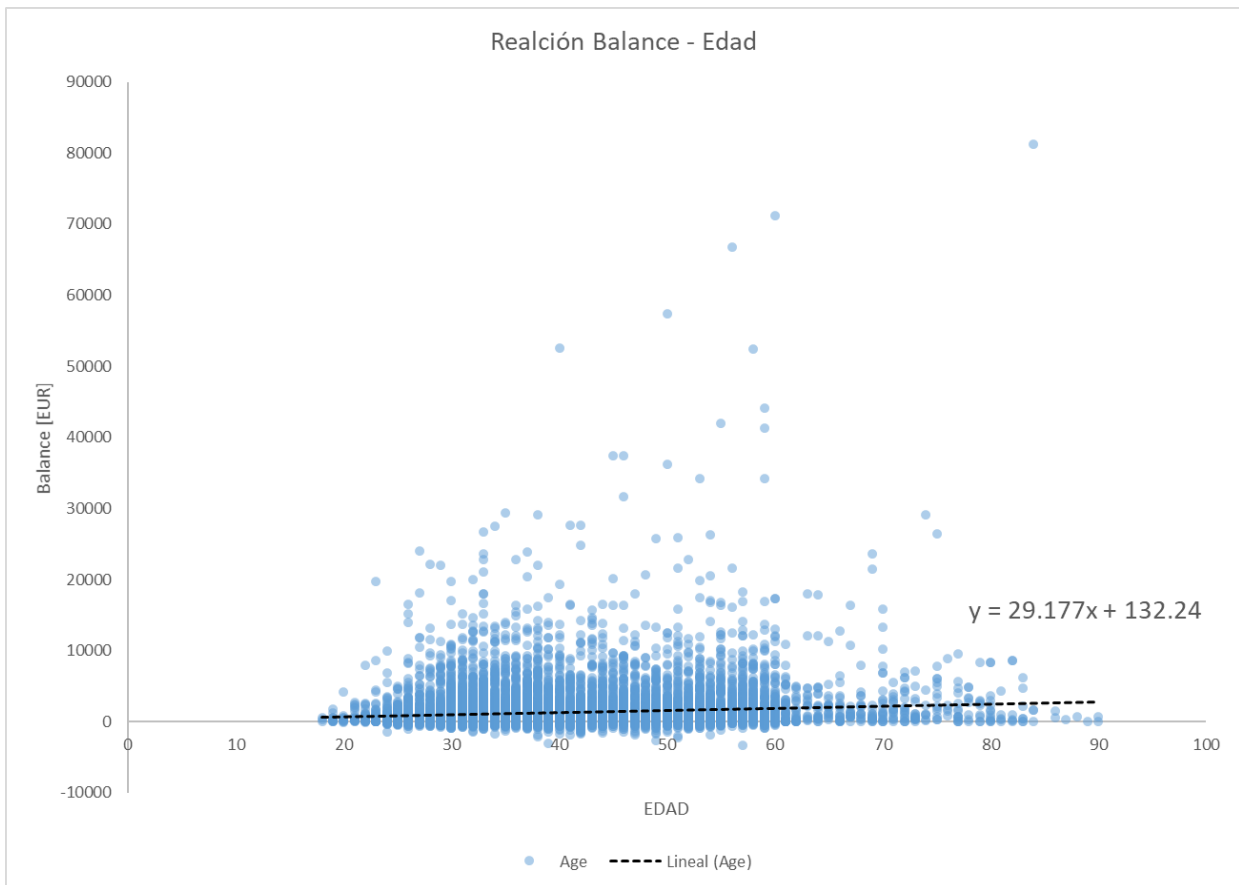
Adicionalmente una significativa diferencia entre la adquisición de préstamos para vivienda y personales respecto a la edad del individuo.

Adicionalmente, se reporta un aumento significativo en el promedio del balance a medida que aumenta el rango de edad de la subpoblación.

Educación	promedio de Balance..euros
80 +	3753.166667
70 - 79	2967.751938
60 - 69	2225.324538
50 - 59	1670.084067
40 - 49	1264.706451
30 - 39	1202.419074
-30	940.070802
Total general	1327.114871

En segunda instancia, se realizan gráficos de dispersión con el objetivo de determinar la relación entre variables de tipo cuantitativo de interés, el análisis de estas en conjunto con el coeficiente de correlación permitirá conocer si existe una influencia significativa entre las variables.

A continuación, se presenta la gráfica de dispersión realizada para las variables Edad y Balance.



La cual permite sospechar de una relación entre dichas variables, sin embargo, deja cabida a pensar que existe influencia de otros factores que afectan la linealidad entre las mismas.

Para el cálculo del coeficiente de correlación se tiene que:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{44668932.66}{\sqrt{(1530948.84)(117058081931.029)}}$$

$$r = 0.105$$

El coeficiente de correlación indica una relación débil entre ambas variables, por lo cual es posible descartar la dependencia lineal entre estas.

