Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

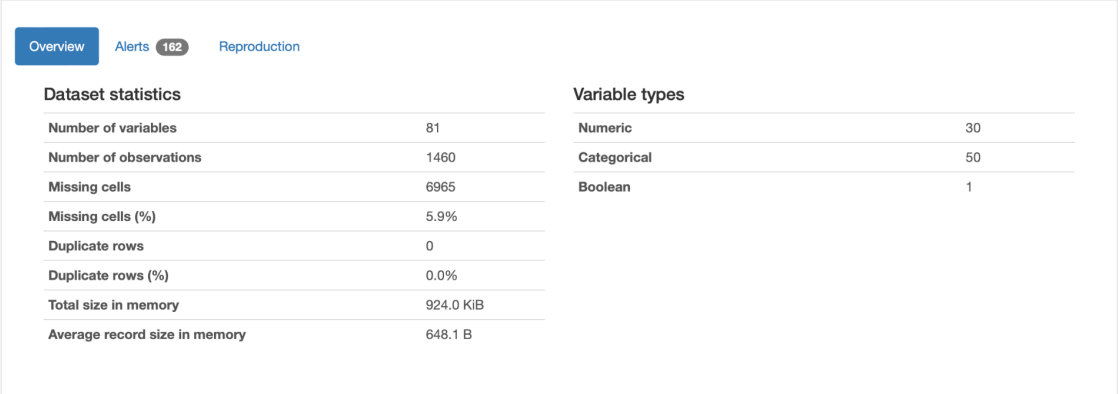
Link al repo:
https://github.com/raulangelj/Laboratorio1An-lisis_Exploratorio_PCA_aprior
Link al drive:
https://docs.google.com/document/d/1v-qLmCEHHOxTXKwv5GC6-Z4n-UJdAz-wKulyPZhPbbc/edit?usp=sharing

Laboratorio 1

Reporte de Datos **17**
Análisis exploratorio 17
Correlaciones 56
Hallazgos y explicación de procedimiento 57
Análisis de componentes principales 58
Análisis factorial 58
Bartlett 58
Hallazgos 59
Matriz de correlación 59
Coeficientes principales. 60
PCA 61
Reglas de asociación 62
Hallazgos y conclusiones 62
Rúbrica 62

#1 Resumen dataset

Overview



#2 Variables a estudiar

Cuantitativa Discreta

Id

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

MSSubClass
LotFrontage
LotArea
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold

Cuantitativa Continua

Cualitativa o Categórica

Alley
Street
LotShape
LandContour
Utilities
LotConfig

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

LandSlope
Neighborhood
Condition1
Condition2
BldgType
HouseStyle
OverallQual
OverallCond
RoofStyle
RoofMatl
Exterior1st
Exterior2nd
MasVnrType
ExterQual
ExterCond
Foundation
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
Heating
HeatingQC
CentralAir
Electrical
KitchenQual
Functional
FireplaceQu
GarageType
GarageFinish
GarageQual
GarageCond
PavedDrive
PoolQC
Fence
MiscFeature
SaleType
SaleCondition

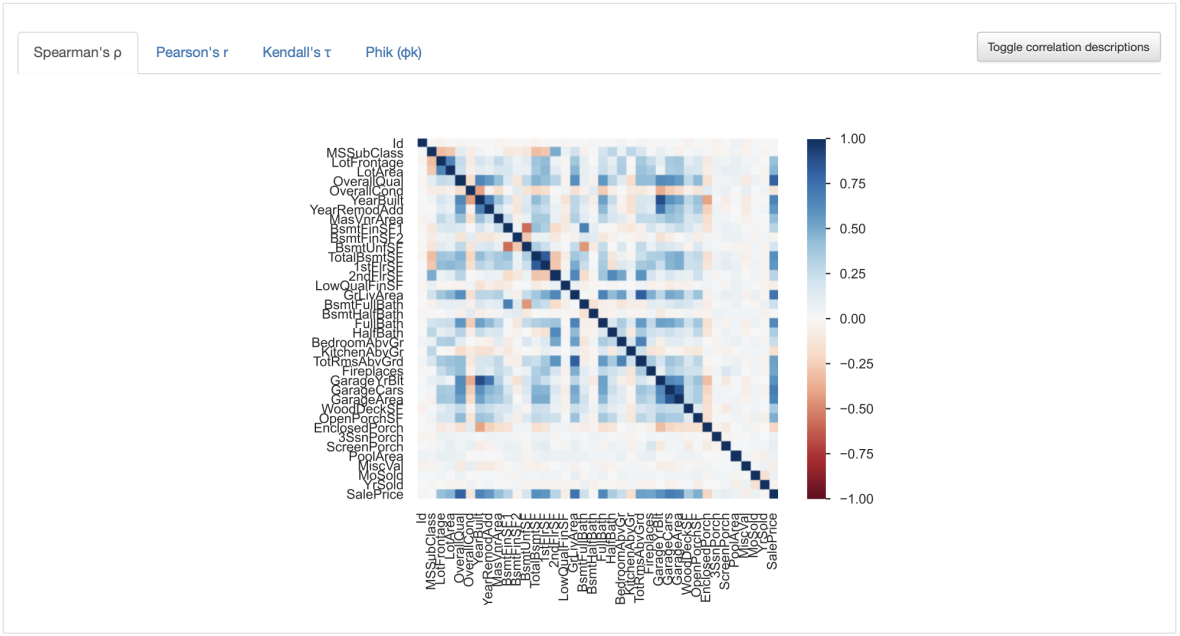
#3 Variables a estudiar

Se pueden observar las gráficas de estos datos en el reporte más abajo.

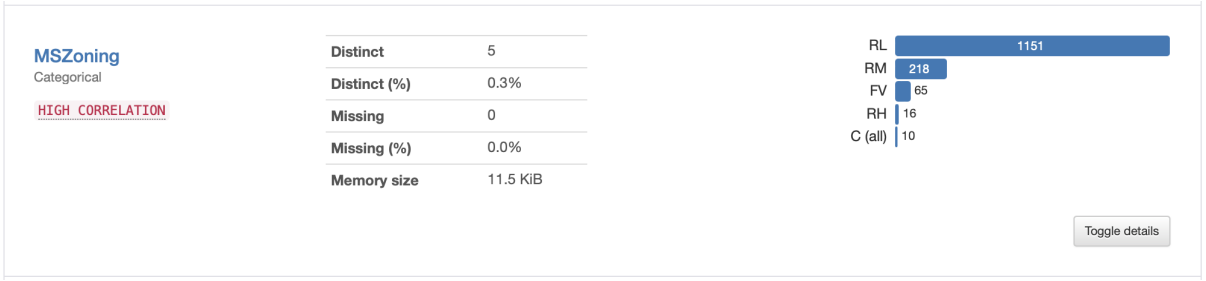
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

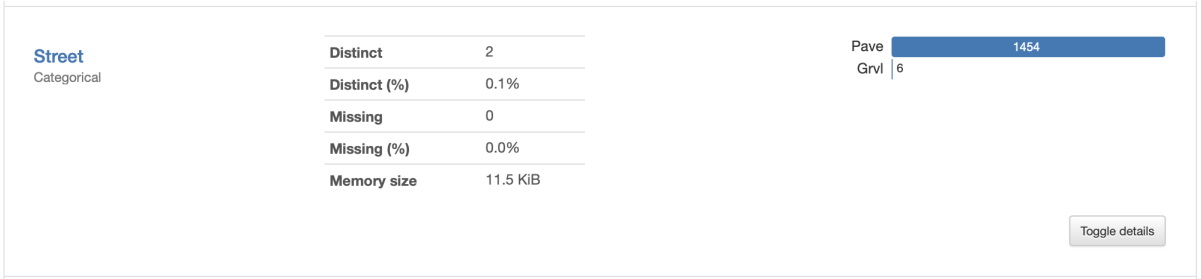
#4 Análisis de correlación, variables numéricas categóricas

Correlations



#5 Variables categóricas análisis



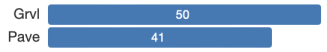


Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Alley
Categorical

HIGH CORRELATION
MISSING

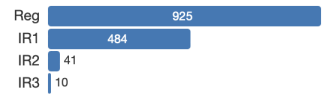
Distinct	2
Distinct (%)	2.2%
Missing	1369
Missing (%)	93.8%
Memory size	11.5 KiB



Toggle details

LotShape
Categorical

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

LandContour
Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Utilities
Categorical

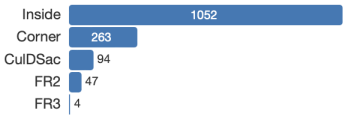
Distinct	2
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

LotConfig
Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

LandSlope
Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

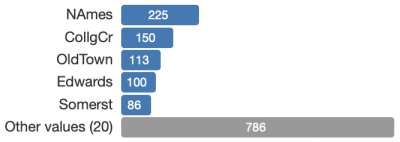
Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Neighborhood
Categorical

HIGH CORRELATION

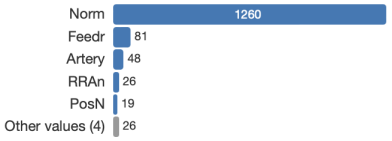
Distinct	25
Distinct (%)	1.7%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Condition1
Categorical

Distinct	9
Distinct (%)	0.6%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Condition2
Categorical

HIGH CORRELATION

Distinct	8
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

BldgType
Categorical

HIGH CORRELATION

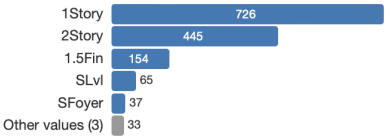
Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

HouseStyle
Categorical

HIGH CORRELATION

Distinct	8
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



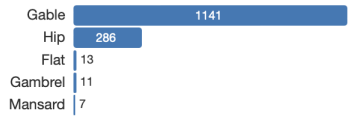
Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

RoofStyle
Categorical

HIGH CORRELATION

Distinct	6
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

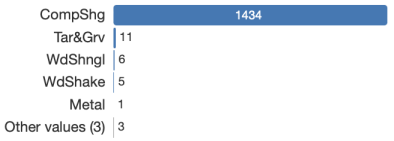


Toggle details

RoofMatl
Categorical

HIGH CORRELATION

Distinct	8
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

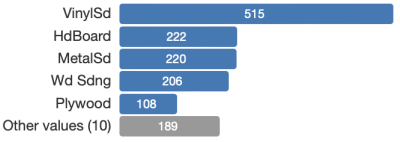


Toggle details

Exterior1st
Categorical

HIGH CORRELATION

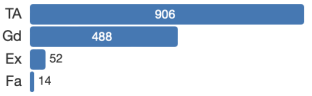
Distinct	15
Distinct (%)	1.0%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



ExterQual
Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

ExterCond
Categorical

HIGH CORRELATION

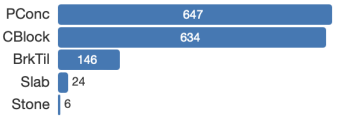
Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Foundation
Categorical

HIGH CORRELATION

Distinct	6
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

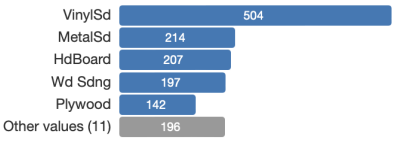


Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Exterior2nd
Categorical

HIGH CORRELATION

Distinct	16
Distinct (%)	1.1%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

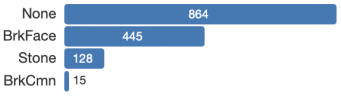


Toggle details

MasVnrType
Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	8
Missing (%)	0.5%
Memory size	11.5 KiB

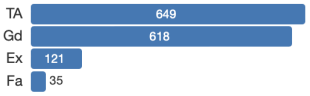


Toggle details

BsmtQual
Categorical

HIGH CORRELATION
MISSING

Distinct	4
Distinct (%)	0.3%
Missing	37
Missing (%)	2.5%
Memory size	11.5 KiB



Toggle details

BsmtCond
Categorical

HIGH CORRELATION
MISSING

Distinct	4
Distinct (%)	0.3%
Missing	37
Missing (%)	2.5%
Memory size	11.5 KiB

Toggle details

BsmtExposure
Categorical

HIGH CORRELATION
MISSING

Distinct	4
Distinct (%)	0.3%
Missing	38
Missing (%)	2.6%
Memory size	11.5 KiB

BsmtFinType2
Categorical

HIGH CORRELATION
MISSING

Distinct	6
Distinct (%)	0.4%
Missing	38
Missing (%)	2.6%
Memory size	11.5 KiB

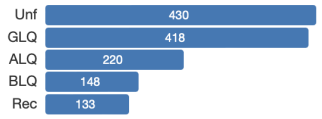
Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

BsmtFinType1
Categorical

HIGH CORRELATION
MISSING

Distinct	6
Distinct (%)	0.4%
Missing	37
Missing (%)	2.5%
Memory size	11.5 KiB



Toggle details

Heating
Categorical

HIGH CORRELATION

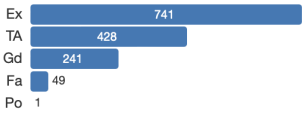
Distinct	6
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

HeatingQC
Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



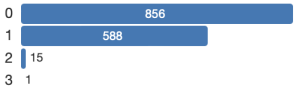
Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

BsmtFullBath
Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

BsmtHalfBath
Categorical

HIGH CORRELATION

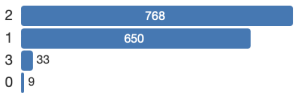
Distinct	3
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

FullBath
Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

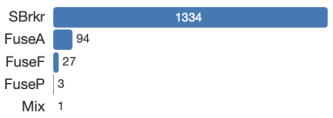


Toggle details

Electrical
Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	1
Missing (%)	0.1%
Memory size	11.5 KiB

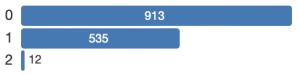


Toggle details

HalfBath
Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	3
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

KitchenAbvGr
Categorical

HIGH CORRELATION

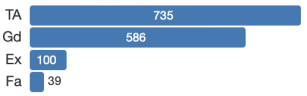
Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

KitchenQual
Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

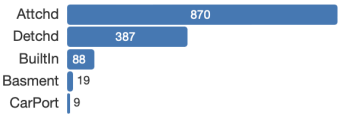


Toggle details

GarageType
Categorical

HIGH CORRELATION
MISSING

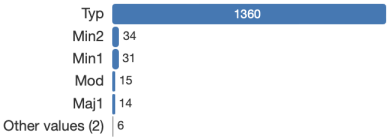
Distinct	6
Distinct (%)	0.4%
Missing	81
Missing (%)	5.5%
Memory size	11.5 KiB



Toggle details

Functional
Categorical

Distinct	7
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

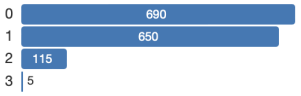


Toggle details

Fireplaces
Categorical

HIGH CORRELATION
HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

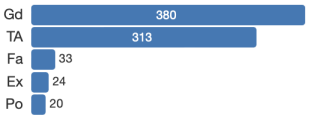


Toggle details

FireplaceQu
Categorical

HIGH CORRELATION
MISSING

Distinct	5
Distinct (%)	0.6%
Missing	690
Missing (%)	47.3%
Memory size	11.5 KiB



Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

GarageQual
Categorical

HIGH CORRELATION
MISSING

Distinct	5
Distinct (%)	0.4%
Missing	81
Missing (%)	5.5%
Memory size	11.5 KiB

Toggle details

GarageCond
Categorical

HIGH CORRELATION
MISSING

Distinct	5
Distinct (%)	0.4%
Missing	81
Missing (%)	5.5%
Memory size	11.5 KiB

Toggle details

PavedDrive
Categorical

HIGH CORRELATION

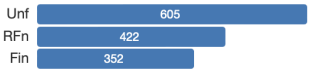
Distinct	3
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

GarageFinish
Categorical

HIGH CORRELATION
MISSING

Distinct	3
Distinct (%)	0.2%
Missing	81
Missing (%)	5.5%
Memory size	11.5 KiB

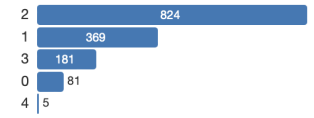


Toggle details

GarageCars
Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

CentralAir
Boolean

HIGH CORRELATION

Distinct	2
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.6 KiB

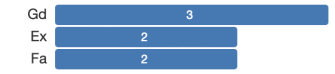
Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

PoolQC
Categorical

HIGH CORRELATION
MISSING

Distinct	3
Distinct (%)	42.9%
Missing	1453
Missing (%)	99.5%
Memory size	11.5 KiB

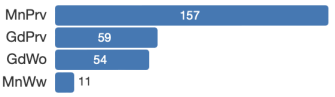


Toggle details

Fence
Categorical

MISSING

Distinct	4
Distinct (%)	1.4%
Missing	1179
Missing (%)	80.8%
Memory size	11.5 KiB



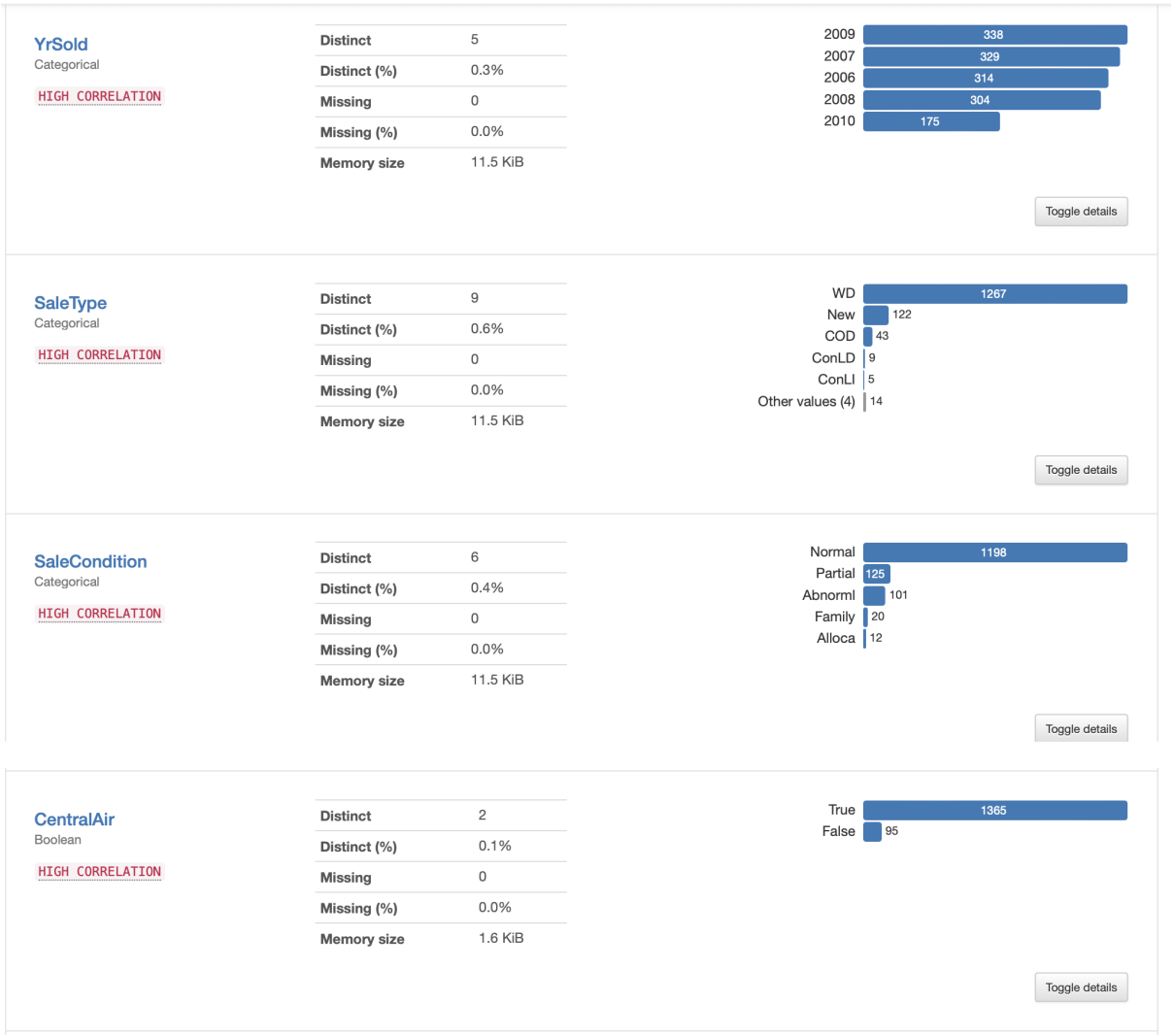
Toggle details

MiscFeature
Categorical

HIGH CORRELATION
MISSING

Distinct	4
Distinct (%)	7.4%
Missing	1406
Missing (%)	96.3%
Memory size	11.5 KiB

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017



#6 Análisis de componentes principales con variables numéricas

Estas gráficas y análisis se pueden observar más detalladamente en el reporte de abajo

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity

chi_square_value,p_value=calculate_bartlett_sphericity(houses_clean)
chi_square_value, p_value
```

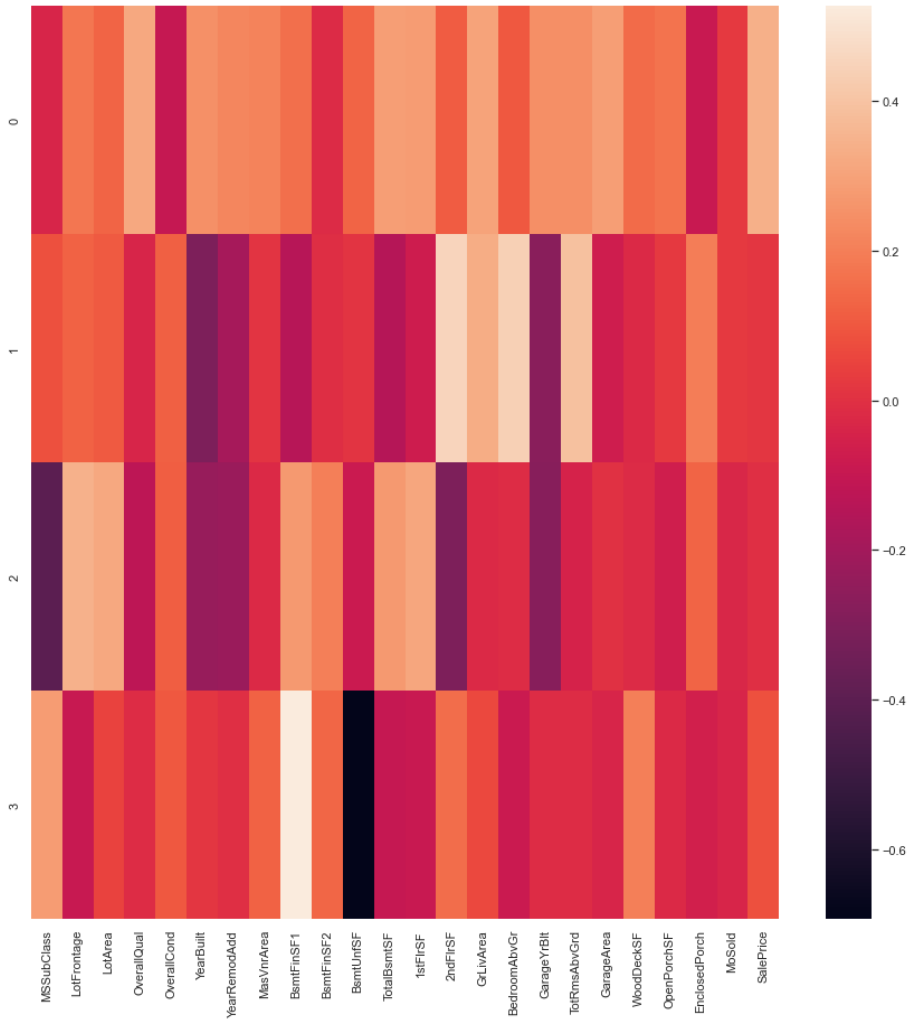
```
(56989.803679788645, 0.0)
```

```
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(houses_clean)
kmo_all, kmo_model
```

```
c:\Users\ALIEWARE\AppData\Local\Programs\Python\Python39\lib\site-packages\factor_analyzer\utils.py:249
  warnings.warn('The inverse of the variance-covariance matrix '

(array([0.67986034, 0.85336964, 0.86511003, 0.91950297, 0.52939275,
        0.76092737, 0.83176101, 0.94283582, 0.70914247, 0.11991146,
        0.64250957, 0.87220836, 0.62952928, 0.45033991, 0.66780668,
        0.77086765, 0.82586225, 0.91502783, 0.91616437, 0.94585033,
        0.93825884, 0.71845741, 0.46740089, 0.92301962]),
 0.7683818262377667)
```

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017



#7 Reglas de asociación interesantes del dataset

```
# El minimo de cobertura o soporte es de 20% y e lminimo de confianza es de 70%
reglas_asociacion = apriori(records, min_support=0.2, min_confidence=0.7)
reglas = list(reglas_asociacion)
```

41]

```
print(len(reglas))
```

42]

```
1531
```

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Reporte de Datos

Análisis exploratorio

Overview

Overview | Alerts 162 | Reproduction

Dataset statistics

Number of variables	81
Number of observations	1460
Missing cells	6965
Missing cells (%)	5.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	924.0 KiB
Average record size in memory	648.1 B

Variable types

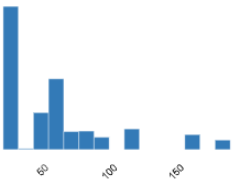
Numeric	30
Categorical	50
Boolean	1

MSSubClass
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	15
Distinct (%)	1.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	56.89726027

Minimum	20
Maximum	190
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	20
5-th percentile	20
Q1	20
median	50
Q3	70
95-th percentile	160
Maximum	190
Range	170
Interquartile range (IQR)	50

Descriptive statistics

Standard deviation	42.30057099
Coefficient of variation (CV)	0.7434553226
Kurtosis	1.580187965
Mean	56.89726027
Median Absolute Deviation (MAD)	30
Skewness	1.407656747
Sum	83070
Variance	1789.338306
Monotonicity	Not monotonic

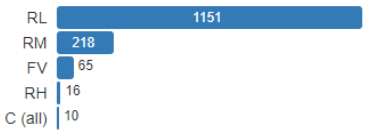
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

MSZoning

Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	2
Mean length	2.034246575
Min length	2

Characters and Unicode

Total characters	2970
Distinct characters	12
Distinct categories	5
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	RL
2nd row	RL
3rd row	RL
4th row	RL
5th row	RL

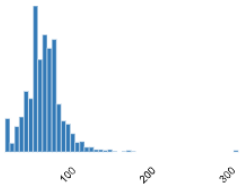
LotFrontage

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	110
Distinct (%)	9.2%
Missing	259
Missing (%)	17.7%
Infinite	0
Infinite (%)	0.0%
Mean	70.04995837

Minimum	21
Maximum	313
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	21
5-th percentile	34
Q1	59
median	69
Q3	80
95-th percentile	107
Maximum	313
Range	292
Interquartile range (IQR)	21

Descriptive statistics

Standard deviation	24.28475177
Coefficient of variation (CV)	0.3466776047
Kurtosis	17.45286726
Mean	70.04995837
Median Absolute Deviation (MAD)	11
Skewness	2.163569142
Sum	84130
Variance	589.7491687
Monotonicity	Not monotonic

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

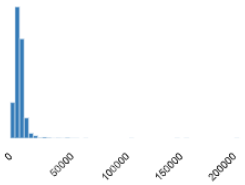
LotArea

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	1073
Distinct (%)	73.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	10516.82808

Minimum	1300
Maximum	215245
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	1300
5-th percentile	3311.7
Q1	7553.5
median	9478.5
Q3	11601.5
95-th percentile	17401.15
Maximum	215245
Range	213945
Interquartile range (IQR)	4048

Descriptive statistics

Standard deviation	9981.264932
Coefficient of variation (CV)	0.949075601
Kurtosis	203.243271
Mean	10516.82808
Median Absolute Deviation (MAD)	1998
Skewness	12.20768785
Sum	15354569
Variance	99625649.65
Monotonicity	Not monotonic

Street

Categorical

Distinct	2
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	4
Median length	4
Mean length	4
Min length	4

Characters and Unicode

Total characters	5840
Distinct characters	7
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Pave
2nd row	Pave
3rd row	Pave
4th row	Pave
5th row	Pave

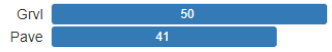
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Alley

Categorical

HIGH CORRELATION
MISSING

Distinct	2
Distinct (%)	2.2%
Missing	1369
Missing (%)	93.8%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	4
Median length	4
Mean length	4
Min length	4

Characters and Unicode

Total characters	364
Distinct characters	7
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

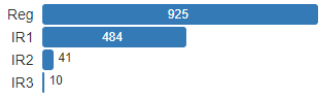
Sample

1st row	Grvl
2nd row	Pave
3rd row	Pave
4th row	Grvl
5th row	Pave

LotShape

Categorical

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	3
Median length	3
Mean length	3
Min length	3

Characters and Unicode

Total characters	4380
Distinct characters	7
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Reg
2nd row	Reg
3rd row	IR1
4th row	IR1
5th row	IR1

LandContour

Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	3
Median length	3
Mean length	3
Min length	3

Characters and Unicode

Total characters	4380
Distinct characters	10
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Lvl
2nd row	Lvl
3rd row	Lvl
4th row	Lvl
5th row	Lvl

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Utilities
Categorical

Distinct	2
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	6
Median length	6
Mean length	6
Min length	6

Characters and Unicode

Total characters	8760
Distinct characters	11
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1
Unique (%)	0.1%

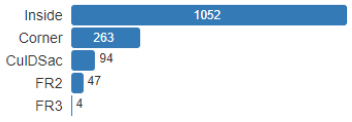
Sample

1st row	AllPub
2nd row	AllPub
3rd row	AllPub
4th row	AllPub
5th row	AllPub

LotConfig
Categorical

HIGH_CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	6
Mean length	5.959589041
Min length	3

Characters and Unicode

Total characters	8701
Distinct characters	19
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Inside
2nd row	FR2
3rd row	Inside
4th row	Corner
5th row	FR2

LandSlope
Categorical

HIGH_CORRELATION

Distinct	3
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	3
Median length	3
Mean length	3
Min length	3

Characters and Unicode

Total characters	4380
Distinct characters	9
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Gtl
2nd row	Gtl
3rd row	Gtl
4th row	Gtl
5th row	Gtl

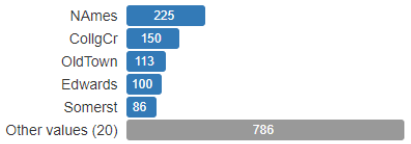
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Neighborhood

Categorical

HIGH_CORRELATION

Distinct	25
Distinct (%)	1.7%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	7
Mean length	6.494520548
Min length	5

Characters and Unicode

Total characters	9482
Distinct characters	38
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

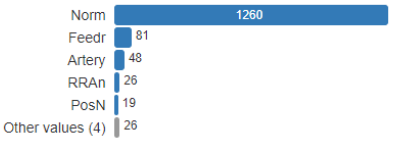
Sample

1st row	CollgCr
2nd row	Veenker
3rd row	CollgCr
4th row	Crawfor
5th row	NoRidge

Condition1

Categorical

Distinct	9
Distinct (%)	0.6%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	6
Median length	4
Mean length	4.121232877
Min length	4

Characters and Unicode

Total characters	6017
Distinct characters	14
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Norm
2nd row	Feedr
3rd row	Norm
4th row	Norm
5th row	Norm

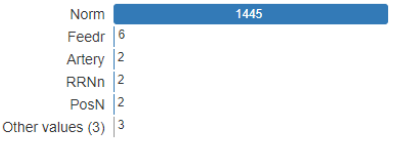
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Condition2

Categorical

HIGH CORRELATION

Distinct	8
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	6
Median length	4
Mean length	4.006849315
Min length	4

Characters and Unicode

Total characters	5850
Distinct characters	14
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	3
Unique (%)	0.2%

Sample

1st row	Norm
2nd row	Norm
3rd row	Norm
4th row	Norm
5th row	Norm

BldgType

Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	6
Median length	4
Mean length	4.299315068
Min length	4

Characters and Unicode

Total characters	6277
Distinct characters	20
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	1Fam
2nd row	1Fam
3rd row	1Fam
4th row	1Fam
5th row	1Fam

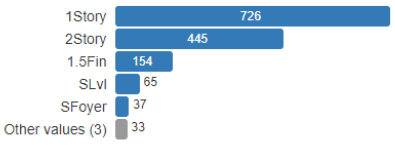
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

HouseStyle

Categorical

HIGH CORRELATION

Distinct	8
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	6
Median length	6
Mean length	5.910958904
Min length	4

Characters and Unicode

Total characters	8630
Distinct characters	18
Distinct categories	4
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	2Story
2nd row	1Story
3rd row	2Story
4th row	2Story
5th row	2Story

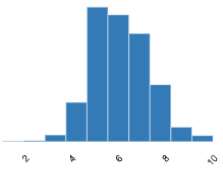
OverallQual

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	10
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	6.099315068

Minimum	1
Maximum	10
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	1
5-th percentile	4
Q1	5
median	6
Q3	7
95-th percentile	8
Maximum	10
Range	9
Interquartile range (IQR)	2

Descriptive statistics

Standard deviation	1.382996547
Coefficient of variation (CV)	0.2267462053
Kurtosis	0.09629277836
Mean	6.099315068
Median Absolute Deviation (MAD)	1
Skewness	0.2169439278
Sum	8905
Variance	1.912679448
Monotonicity	Not monotonic

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

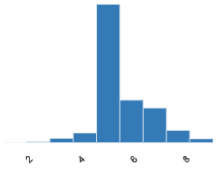
OverallCond

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	9
Distinct (%)	0.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5.575342466

Minimum	1
Maximum	9
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	1
5-th percentile	4
Q1	5
median	5
Q3	6
95-th percentile	8
Maximum	9
Range	8
Interquartile range (IQR)	1

Descriptive statistics

Standard deviation	1.112799337
Coefficient of variation (CV)	0.1995930014
Kurtosis	1.106413461
Mean	5.575342466
Median Absolute Deviation (MAD)	0
Skewness	0.6930674725
Sum	8140
Variance	1.238322364
Monotonicity	Not monotonic

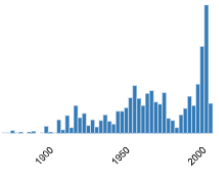
YearBuilt

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	112
Distinct (%)	7.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1971.267808

Minimum	1872
Maximum	2010
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	1872
5-th percentile	1916
Q1	1954
median	1973
Q3	2000
95-th percentile	2007
Maximum	2010
Range	138
Interquartile range (IQR)	46

Descriptive statistics

Standard deviation	30.20290404
Coefficient of variation (CV)	0.01532156307
Kurtosis	-0.4395519416
Mean	1971.267808
Median Absolute Deviation (MAD)	25
Skewness	-0.6134611725
Sum	2878051
Variance	912.2154126
Monotonicity	Not monotonic

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

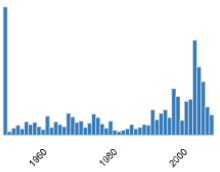
YearRemodAdd

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	61
Distinct (%)	4.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1984.865753

Minimum	1950
Maximum	2010
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	1950
5-th percentile	1950
Q1	1967
median	1994
Q3	2004
95-th percentile	2007
Maximum	2010
Range	60
Interquartile range (IQR)	37

Descriptive statistics

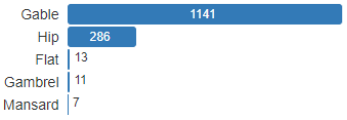
Standard deviation	20.64540681
Coefficient of variation (CV)	0.01040141217
Kurtosis	-1.272245192
Mean	1984.865753
Median Absolute Deviation (MAD)	13
Skewness	-0.5035620027
Sum	2897904
Variance	426.2328223
Monotonicity	Not monotonic

RoofStyle

Categorical

HIGH CORRELATION

Distinct	6
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	5
Mean length	4.62260274
Min length	3

Characters and Unicode

Total characters	6749
Distinct characters	18
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Gable
2nd row	Gable
3rd row	Gable
4th row	Gable
5th row	Gable

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

RoofMatl

Categorical

HIGH CORRELATION

Distinct	8
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	7
Mean length	6.996575342
Min length	4

Characters and Unicode

Total characters	10215
Distinct characters	25
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	4
Unique (%)	0.3%

Sample

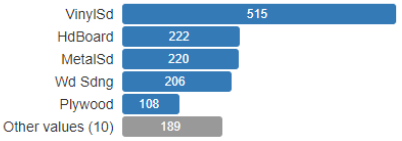
1st row	CompShg
2nd row	CompShg
3rd row	CompShg
4th row	CompShg
5th row	CompShg

Exterior1st

Categorical

HIGH CORRELATION

Distinct	15
Distinct (%)	1.0%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	7
Mean length	6.979452055
Min length	5

Characters and Unicode

Total characters	10190
Distinct characters	32
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	3
Unique (%)	0.2%

Sample

1st row	VinylSd
2nd row	MetalSd
3rd row	VinylSd
4th row	Wd Sdng
5th row	VinylSd

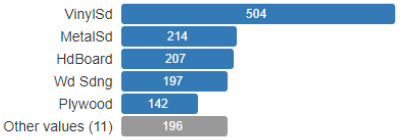
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Exterior2nd

Categorical

HIGH_CORRELATION

Distinct	16
Distinct (%)	1.1%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	7
Mean length	6.973287671
Min length	5

Characters and Unicode

Total characters	10181
Distinct characters	33
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	2
Unique (%)	0.1%

Sample

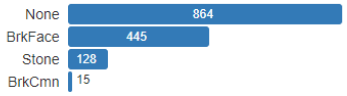
1st row	VinylSd
2nd row	MetalSd
3rd row	VinylSd
4th row	Wd Shng
5th row	VinylSd

MasVnrType

Categorical

HIGH_CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	8
Missing (%)	0.5%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	4
Mean length	5.028236915
Min length	4

Characters and Unicode

Total characters	7301
Distinct characters	14
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	BrkFace
2nd row	None
3rd row	BrkFace
4th row	None
5th row	BrkFace

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

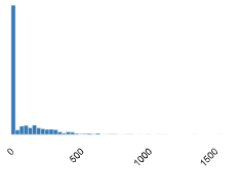
MasVnrArea

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
ZEROS

Distinct	327
Distinct (%)	22.5%
Missing	8
Missing (%)	0.5%
Infinite	0
Infinite (%)	0.0%
Mean	103.6852617

Minimum	0
Maximum	1600
Zeros	861
Zeros (%)	59.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	166
95-th percentile	456
Maximum	1600
Range	1600
Interquartile range (IQR)	166

Descriptive statistics

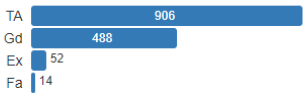
Standard deviation	181.0662066
Coefficient of variation (CV)	1.746306115
Kurtosis	10.08241732
Mean	103.6852617
Median Absolute Deviation (MAD)	0
Skewness	2.66908421
Sum	150551
Variance	32784.97117
Monotonicity	Not monotonic

ExterQual

Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	2920
Distinct characters	8
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Gd
2nd row	TA
3rd row	Gd
4th row	TA
5th row	Gd

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

ExterCond

Categorical

HIGH_CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	2920
Distinct characters	10
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1
Unique (%)	0.1%

Sample

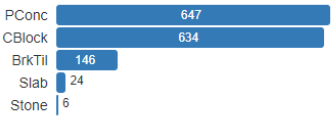
1st row	TA
2nd row	TA
3rd row	TA
4th row	TA
5th row	TA

Foundation

Categorical

HIGH_CORRELATION

Distinct	6
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	6
Median length	6
Mean length	5.515753425
Min length	4

Characters and Unicode

Total characters	8053
Distinct characters	18
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	PConc
2nd row	CBlock
3rd row	PConc
4th row	BrkTil
5th row	PConc

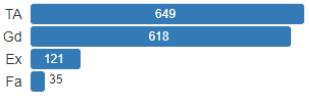
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

BsmtQual

Categorical

HIGH CORRELATION
MISSING

Distinct	4
Distinct (%)	0.3%
Missing	37
Missing (%)	2.5%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	2846
Distinct characters	8
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Gd
2nd row	Gd
3rd row	Gd
4th row	TA
5th row	Gd

BsmtCond

Categorical

HIGH CORRELATION
MISSING

Distinct	4
Distinct (%)	0.3%
Missing	37
Missing (%)	2.5%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	2846
Distinct characters	8
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	TA
2nd row	TA
3rd row	TA
4th row	Gd
5th row	TA

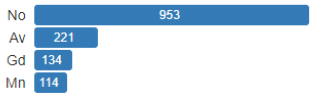
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

BsmtExposure

Categorical

HIGH CORRELATION
MISSING

Distinct	4
Distinct (%)	0.3%
Missing	38
Missing (%)	2.6%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	2844
Distinct characters	8
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

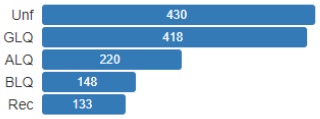
1st row	No
2nd row	Gd
3rd row	Mn
4th row	No
5th row	Av

BsmtFinType1

Categorical

HIGH CORRELATION
MISSING

Distinct	6
Distinct (%)	0.4%
Missing	37
Missing (%)	2.5%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	3
Median length	3
Mean length	3
Min length	3

Characters and Unicode

Total characters	4269
Distinct characters	12
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	GLQ
2nd row	ALQ
3rd row	GLQ
4th row	ALQ
5th row	GLQ

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

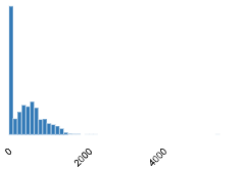
BsmtFinSF1

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	637
Distinct (%)	43.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	443.639726

Minimum	0
Maximum	5644
Zeros	467
Zeros (%)	32.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	383.5
Q3	712.25
95-th percentile	1274
Maximum	5644
Range	5644
Interquartile range (IQR)	712.25

Descriptive statistics

Standard deviation	456.0980908
Coefficient of variation (CV)	1.028082167
Kurtosis	11.11823629
Mean	443.639726
Median Absolute Deviation (MAD)	383.5
Skewness	1.685503072
Sum	647714
Variance	208025.4685
Monotonicity	Not monotonic

BsmtFinType2

Categorical

HIGH CORRELATION
MISSING

Distinct	6
Distinct (%)	0.4%
Missing	38
Missing (%)	2.6%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	3
Median length	3
Mean length	3
Min length	3

Characters and Unicode

Total characters	4266
Distinct characters	12
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Unf
2nd row	Unf
3rd row	Unf
4th row	Unf
5th row	Unf

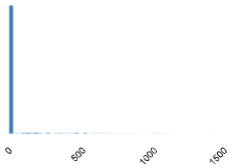
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

BsmtFinSF2

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
ZEROS

Distinct	144	Minimum	0
Distinct (%)	9.9%	Maximum	1474
Missing	0	Zeros	1293
Missing (%)	0.0%	Zeros (%)	88.6%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	46.54931507	Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

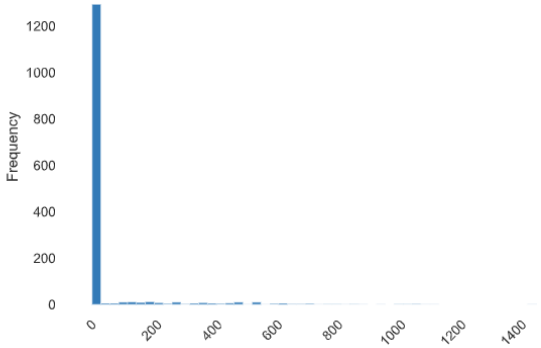
Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	0
95-th percentile	396.2
Maximum	1474
Range	1474
Interquartile range (IQR)	0

Descriptive statistics

Standard deviation	161.3192728
Coefficient of variation (CV)	3.465556315
Kurtosis	20.11333755
Mean	46.54931507
Median Absolute Deviation (MAD)	0
Skewness	4.255261109
Sum	67962
Variance	26023.90778
Monotonicity	Not monotonic

Statistics | Histogram | Common values | Extreme values



Histogram with fixed size bins (bins=50)

Statistics | Histogram | Common values | Extreme values

Value	Count	Frequency (%)
0	1293	88.6%
180	5	0.3%
374	3	0.2%
551	2	0.1%
147	2	0.1%
294	2	0.1%
391	2	0.1%
539	2	0.1%
96	2	0.1%
480	2	0.1%
Other values (134)	145	9.9%

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

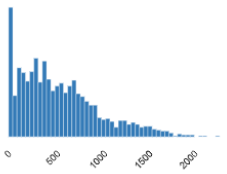
BsmtUnfSF

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	780
Distinct (%)	53.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	567.240411

Minimum	0
Maximum	2336
Zeros	118
Zeros (%)	8.1%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	223
median	477.5
Q3	808
95-th percentile	1468
Maximum	2336
Range	2336
Interquartile range (IQR)	585

Descriptive statistics

Standard deviation	441.8669553
Coefficient of variation (CV)	0.7789765094
Kurtosis	0.4749939878
Mean	567.240411
Median Absolute Deviation (MAD)	288
Skewness	0.9202684528
Sum	828171
Variance	195246.4062
Monotonicity	Not monotonic

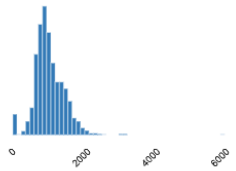
TotalBsmtSF

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	721
Distinct (%)	49.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1057.429452

Minimum	0
Maximum	6110
Zeros	37
Zeros (%)	2.5%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	519.3
Q1	795.75
median	991.5
Q3	1298.25
95-th percentile	1753
Maximum	6110
Range	6110
Interquartile range (IQR)	502.5

Descriptive statistics

Standard deviation	438.7053245
Coefficient of variation (CV)	0.4148790481
Kurtosis	13.25048328
Mean	1057.429452
Median Absolute Deviation (MAD)	234.5
Skewness	1.524254549
Sum	1543847
Variance	192462.3617
Monotonicity	Not monotonic

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Heating

Categorical

HIGH CORRELATION

Distinct	6
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	5
Median length	4
Mean length	4.000684932
Min length	4

Characters and Unicode

Total characters	5841
Distinct characters	13
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1
Unique (%)	0.1%

Sample

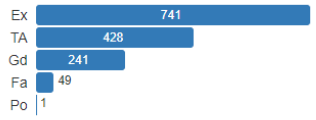
1st row	GasA
2nd row	GasA
3rd row	GasA
4th row	GasA
5th row	GasA

HeatingQC

Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	2920
Distinct characters	10
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1
Unique (%)	0.1%

Sample

1st row	Ex
2nd row	Ex
3rd row	Ex
4th row	Gd
5th row	Ex

CentralAir

Boolean

HIGH CORRELATION

Distinct	2
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.6 KiB

Toggle details

Common Values | Category Frequency Plot

Value	Count	Frequency (%)
True	1365	93.5%
False	95	6.5%

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Electrical

Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	1
Missing (%)	0.1%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	5
Median length	5
Mean length	4.998629198
Min length	3

Characters and Unicode

Total characters	7293
Distinct characters	13
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character

Unique

Unique	1
Unique (%)	0.1%

Sample

1st row	SBrkr
2nd row	SBrkr
3rd row	SBrkr
4th row	SBrkr
5th row	SBrkr

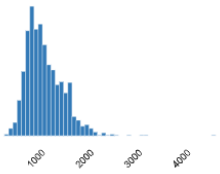
1stFlrSF

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	753
Distinct (%)	51.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1162.626712

Minimum	334
Maximum	4692
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	334
5-th percentile	672.95
Q1	882
median	1087
Q3	1391.25
95-th percentile	1831.25
Maximum	4692
Range	4358
Interquartile range (IQR)	509.25

Descriptive statistics

Standard deviation	386.587738
Coefficient of variation (CV)	0.3325123481
Kurtosis	5.745841482
Mean	1162.626712
Median Absolute Deviation (MAD)	234.5
Skewness	1.376756622
Sum	1697435
Variance	149450.0792
Monotonicity	Not monotonic

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

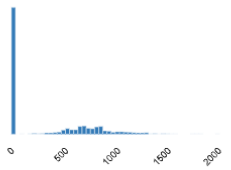
2ndFlrSF

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	417
Distinct (%)	28.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	346.9924658

Minimum	0
Maximum	2065
Zeros	829
Zeros (%)	56.8%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

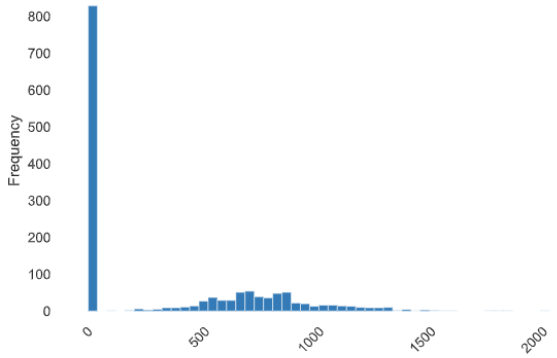
Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	728
95-th percentile	1141.05
Maximum	2065
Range	2065
Interquartile range (IQR)	728

Descriptive statistics

Standard deviation	436.5284359
Coefficient of variation (CV)	1.258034335
Kurtosis	-0.5534635576
Mean	346.9924658
Median Absolute Deviation (MAD)	0
Skewness	0.8130298163
Sum	506609
Variance	190557.0753
Monotonicity	Not monotonic

Statistics | Histogram | Common values | Extreme values



Histogram with fixed size bins (bins=50)

Statistics | Histogram | Common values | Extreme values

Value	Count	Frequency (%)
0	829	56.8%
728	10	0.7%
504	9	0.6%
546	8	0.5%
672	8	0.5%
600	7	0.5%
720	7	0.5%
896	6	0.4%
862	5	0.3%
780	5	0.3%
Other values (407)	566	38.8%

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

LowQualFinSF

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
ZEROS

Distinct	24
Distinct (%)	1.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5.844520548

Minimum	0
Maximum	572
Zeros	1434
Zeros (%)	98.2%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB

Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

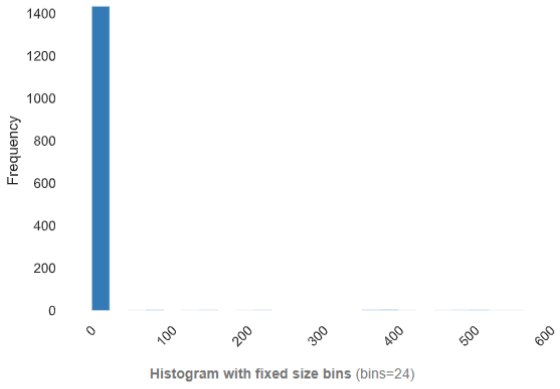
Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	0
95-th percentile	0
Maximum	572
Range	572
Interquartile range (IQR)	0

Descriptive statistics

Standard deviation	48.62308143
Coefficient of variation (CV)	8.319430317
Kurtosis	83.23481667
Mean	5.844520548
Median Absolute Deviation (MAD)	0
Skewness	9.011341288
Sum	8533
Variance	2364.204048
Monotonicity	Not monotonic

Toggle details

Statistics | Histogram | Common values | Extreme values



Histogram with fixed size bins (bins=24)

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

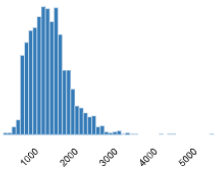
Statistics | Histogram | Common values | Extreme values

Value	Count	Frequency (%)
0	1434	98.2%
80	3	0.2%
360	2	0.1%
205	1	0.1%
479	1	0.1%
397	1	0.1%
514	1	0.1%
120	1	0.1%
481	1	0.1%
232	1	0.1%
Other values (14)	14	1.0%

GrLivArea
Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	861	**Minimum**	334
Distinct (%)	59.0%	**Maximum**	5642
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1515.463699	**Memory size**	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	334
5-th percentile	848
Q1	1129.5
median	1464
Q3	1776.75
95-th percentile	2466.1
Maximum	5642
Range	5308
Interquartile range (IQR)	647.25

Descriptive statistics

Standard deviation	525.4803834
Coefficient of variation (CV)	0.3467456092
Kurtosis	4.895120581
Mean	1515.463699
Median Absolute Deviation (MAD)	326
Skewness	1.366560356
Sum	2212577
Variance	276129.6334
Monotonicity	Not monotonic

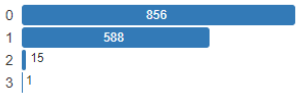
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

BsmtFullBath

Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1460
Distinct characters	4
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1
Unique (%)	0.1%

Sample

1st row	1
2nd row	0
3rd row	1
4th row	1
5th row	1

BsmtHalfBath

Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1460
Distinct characters	3
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	0
2nd row	1
3rd row	0
4th row	0
5th row	0

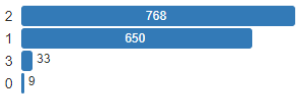
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

FullBath

Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1460
Distinct characters	4
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

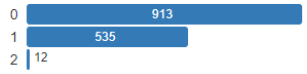
1st row	2
2nd row	2
3rd row	2
4th row	1
5th row	2

HalfBath

Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	3
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1460
Distinct characters	3
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	1
2nd row	0
3rd row	1
4th row	0
5th row	1

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

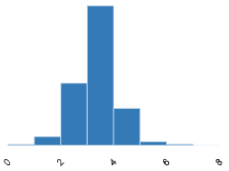
BedroomAbvGr

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	8
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.866438356

Minimum	0
Maximum	8
Zeros	6
Zeros (%)	0.4%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	2
Q1	2
median	3
Q3	3
95-th percentile	4
Maximum	8
Range	8
Interquartile range (IQR)	1

Descriptive statistics

Standard deviation	0.8157780441
Coefficient of variation (CV)	0.2845964025
Kurtosis	2.230874582
Mean	2.866438356
Median Absolute Deviation (MAD)	0
Skewness	0.2117900963
Sum	4185
Variance	0.6654938173
Monotonicity	Not monotonic

KitchenAbvGr

Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1460
Distinct characters	4
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1
Unique (%)	0.1%

Sample

1st row	1
2nd row	1
3rd row	1
4th row	1
5th row	1

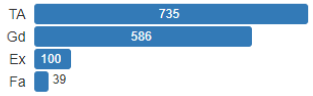
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

KitchenQual

Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	2920
Distinct characters	8
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Gd
2nd row	TA
3rd row	Gd
4th row	Gd
5th row	Gd

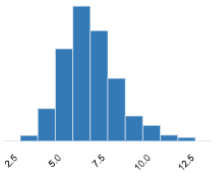
TotRmsAbvGrd

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	12
Distinct (%)	0.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	6.517808219

Minimum	2
Maximum	14
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	2
5-th percentile	4
Q1	5
median	6
Q3	7
95-th percentile	10
Maximum	14
Range	12
Interquartile range (IQR)	2

Descriptive statistics

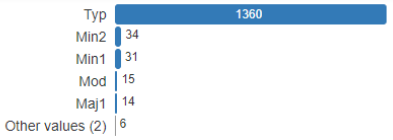
Standard deviation	1.625393291
Coefficient of variation (CV)	0.2493772808
Kurtosis	0.8807615657
Mean	6.517808219
Median Absolute Deviation (MAD)	1
Skewness	0.6763408364
Sum	9516
Variance	2.641903349
Monotonicity	Not monotonic

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Functional

Categorical

Distinct	7
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	4
Median length	3
Mean length	3.057534247
Min length	3

Characters and Unicode

Total characters	4464
Distinct characters	15
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1
Unique (%)	0.1%

Sample

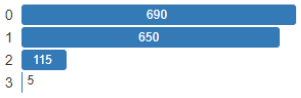
1st row	Typ
2nd row	Typ
3rd row	Typ
4th row	Typ
5th row	Typ

Fireplaces

Categorical

HIGH_CORRELATION
HIGH_CORRELATION

Distinct	4
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1460
Distinct characters	4
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

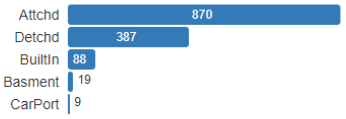
1st row	0
2nd row	1
3rd row	1
4th row	1
5th row	1

GarageType

Categorical

HIGH_CORRELATION
MISSING

Distinct	6
Distinct (%)	0.4%
Missing	81
Missing (%)	5.5%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	7
Median length	6
Mean length	6.084118927
Min length	6

Characters and Unicode

Total characters	8390
Distinct characters	24
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Attchd
2nd row	Attchd
3rd row	Attchd
4th row	Detchd
5th row	Attchd

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

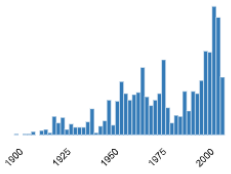
GarageYrBlt

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	97
Distinct (%)	7.0%
Missing	81
Missing (%)	5.5%
Infinite	0
Infinite (%)	0.0%
Mean	1978.506164

Minimum	1900
Maximum	2010
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

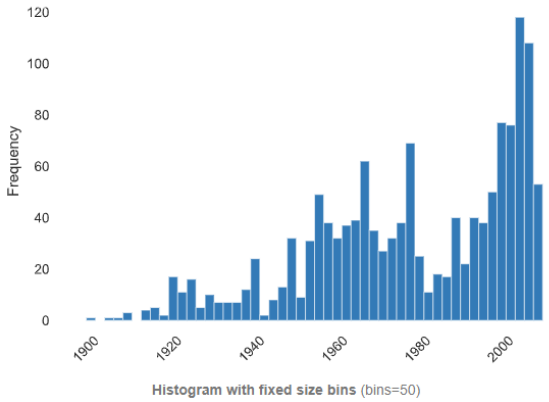
Minimum	1900
5-th percentile	1930
Q1	1961
median	1980
Q3	2002
95-th percentile	2007
Maximum	2010
Range	110
Interquartile range (IQR)	41

Descriptive statistics

Standard deviation	24.68972477
Coefficient of variation (CV)	0.01247897288
Kurtosis	-0.418340998
Mean	1978.506164
Median Absolute Deviation (MAD)	21
Skewness	-0.6494146239
Sum	2728360
Variance	609.5825091
Monotonicity	Not monotonic

Toggle details

Statistics | Histogram | Common values | Extreme values



Histogram with fixed size bins (bins=50)

Statistics | Histogram | Common values | Extreme values

Value	Count	Frequency (%)
2005	65	4.5%
2006	59	4.0%
2004	53	3.6%
2003	50	3.4%
2007	49	3.4%
1977	35	2.4%
1998	31	2.1%
1999	30	2.1%
1976	29	2.0%
2008	29	2.0%
Other values (87)	949	65.0%
(Missing)	81	5.5%

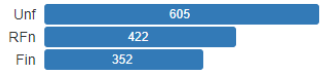
Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

GarageFinish

Categorical

HIGH CORRELATION
MISSING

Distinct	3
Distinct (%)	0.2%
Missing	81
Missing (%)	5.5%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	3
Median length	3
Mean length	3
Min length	3

Characters and Unicode

Total characters	4137
Distinct characters	6
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

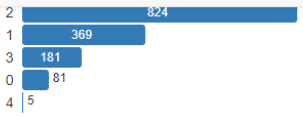
1st row	RFn
2nd row	RFn
3rd row	RFn
4th row	Unf
5th row	RFn

GarageCars

Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

Overview | Categories | Words | Characters

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1460
Distinct characters	5
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	2
2nd row	2
3rd row	2
4th row	3
5th row	3

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

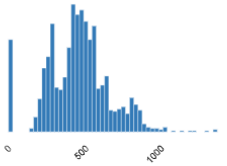
GarageArea

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	441
Distinct (%)	30.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	472.980137

Minimum	0
Maximum	1418
Zeros	81
Zeros (%)	5.5%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

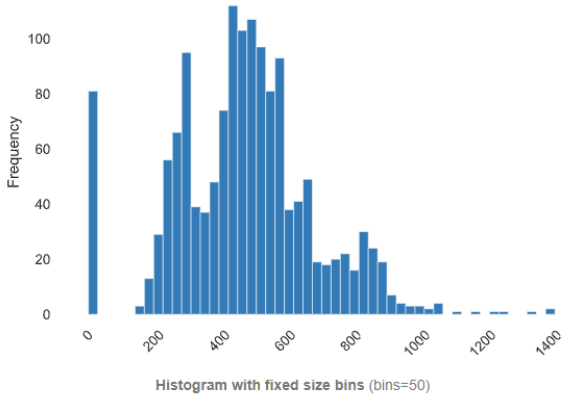
Quantile statistics

Minimum	0
5-th percentile	0
Q1	334.5
median	480
Q3	576
95-th percentile	850.1
Maximum	1418
Range	1418
Interquartile range (IQR)	241.5

Descriptive statistics

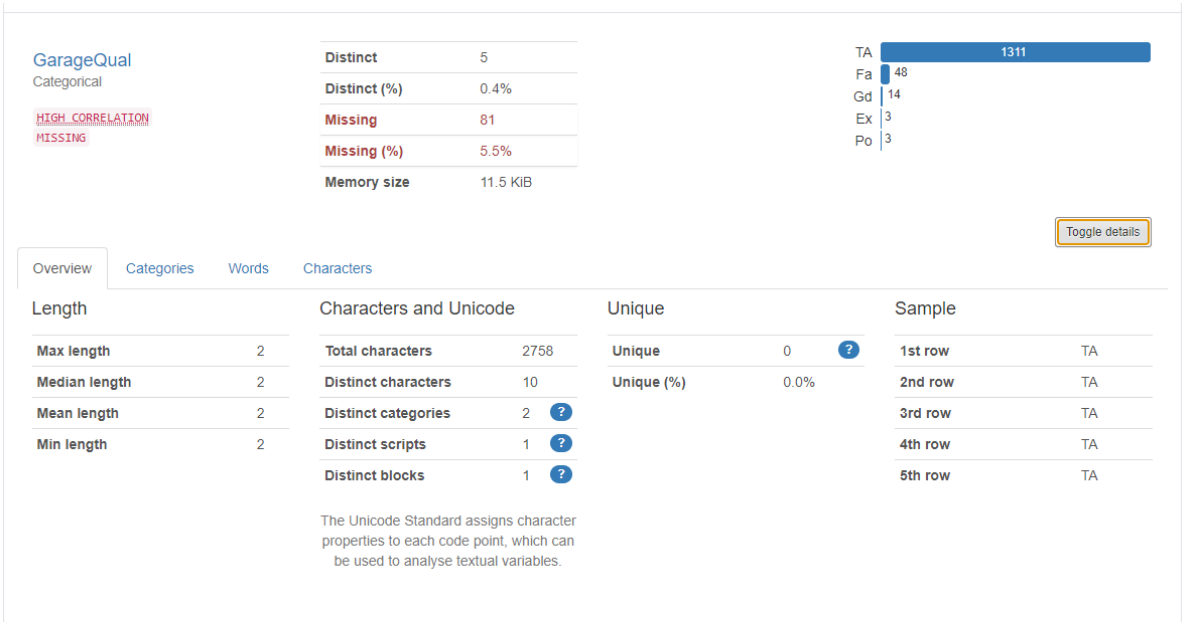
Standard deviation	213.8048415
Coefficient of variation (CV)	0.452037675
Kurtosis	0.9170672023
Mean	472.980137
Median Absolute Deviation (MAD)	120
Skewness	0.1799809067
Sum	690551
Variance	45712.51023
Monotonicity	Not monotonic

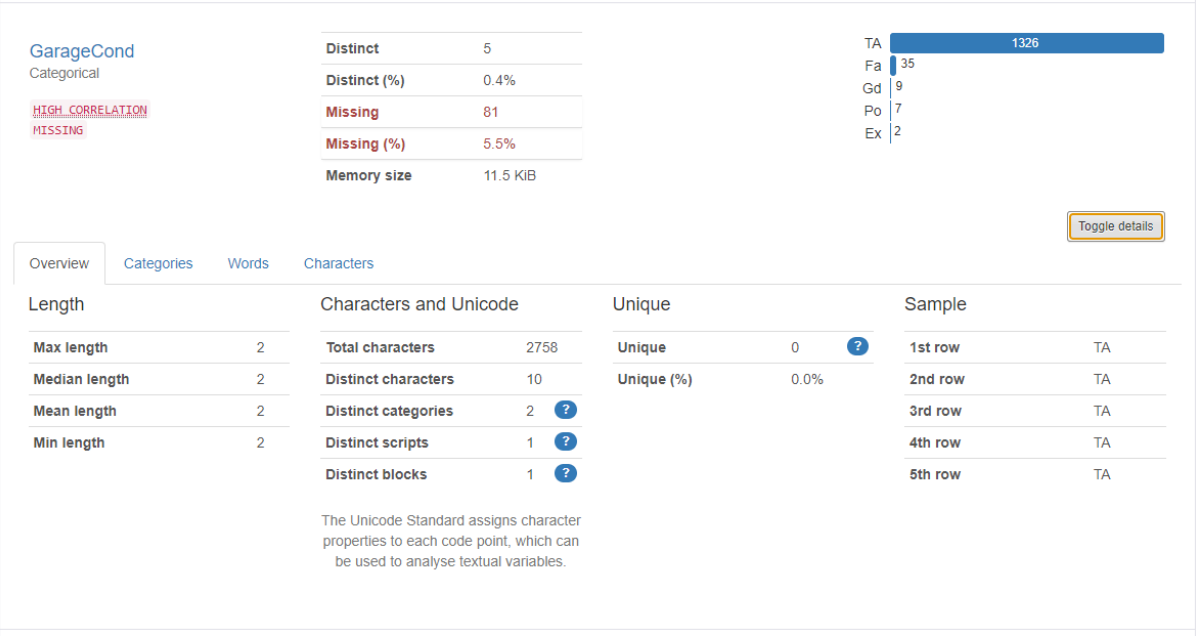
Statistics | Histogram | Common values | Extreme values



Histogram with fixed size bins (bins=50)

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017





Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

PavedDrive

Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Overview | Categories | Words | Characters

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1460
Distinct characters	3
Distinct categories	1
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Y
2nd row	Y
3rd row	Y
4th row	Y
5th row	Y

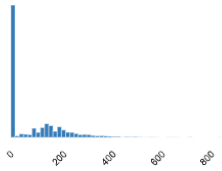
WoodDeckSF

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
ZEROS

Distinct	274
Distinct (%)	18.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	94.24452055

Minimum	0
Maximum	857
Zeros	761
Zeros (%)	52.1%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

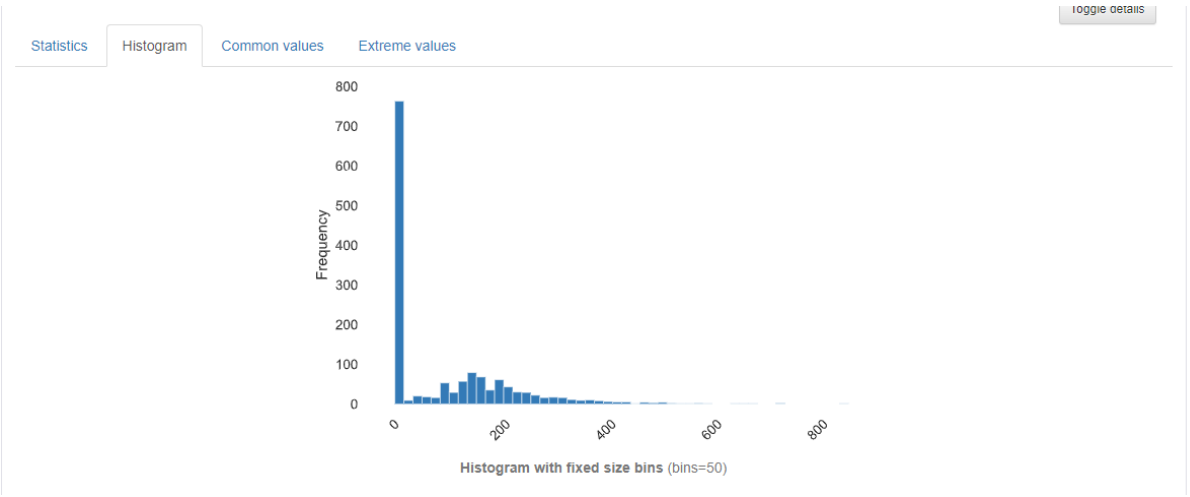
Quantile statistics

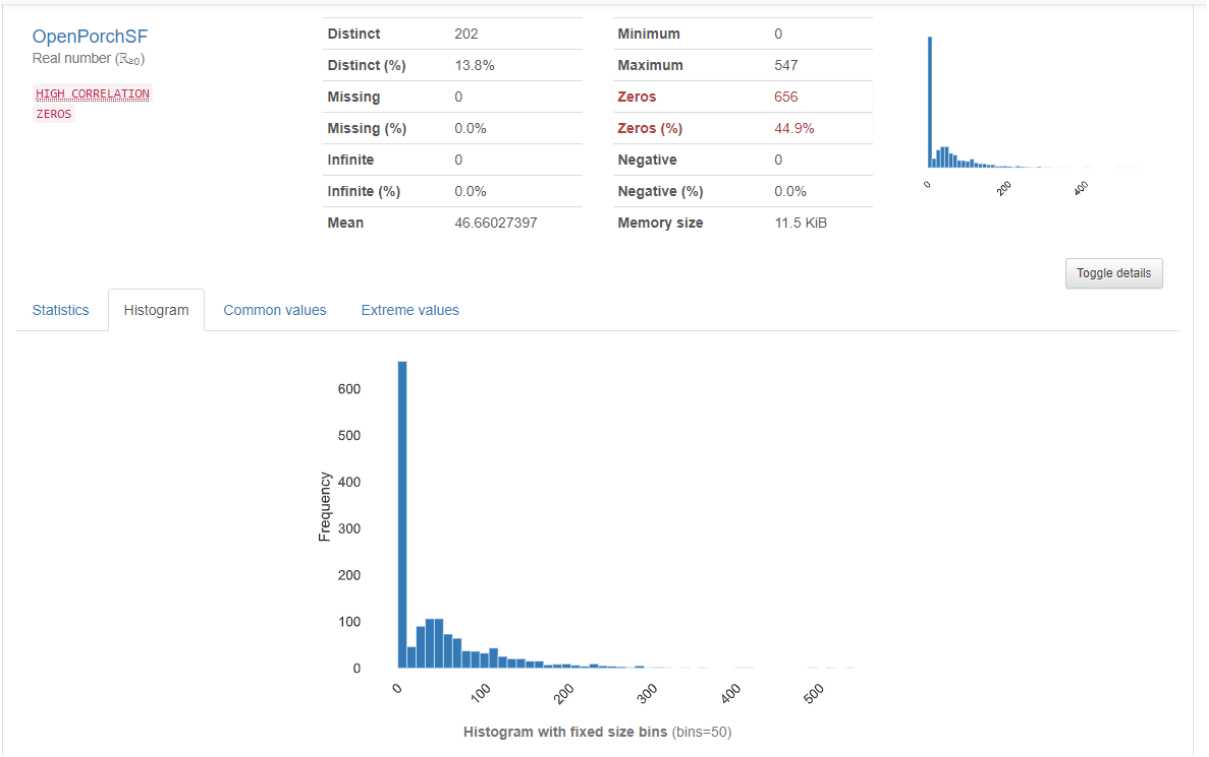
Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	168
95-th percentile	335
Maximum	857
Range	857
Interquartile range (IQR)	168

Descriptive statistics

Standard deviation	125.3387944
Coefficient of variation (CV)	1.329931901
Kurtosis	2.992950925
Mean	94.24452055
Median Absolute Deviation (MAD)	0
Skewness	1.541375757
Sum	137597
Variance	15709.81337
Monotonicity	Not monotonic

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017





Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

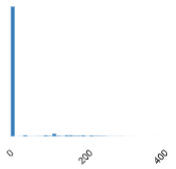
EnclosedPorch

Real number (ℝ≥0)

HIGH CORRELATION
ZEROS

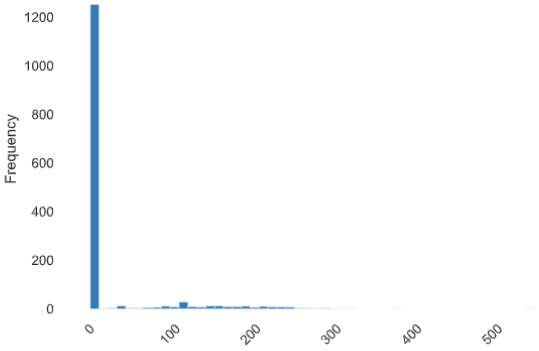
Distinct	120
Distinct (%)	8.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	21.95410959

Minimum	0
Maximum	552
Zeros	1252
Zeros (%)	85.8%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values



Histogram with fixed size bins (bins=50)

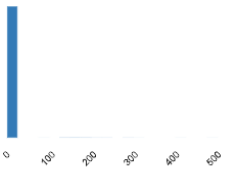
3SsnPorch

Real number (ℝ≥0)

ZEROS

Distinct	20
Distinct (%)	1.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	3.409589041

Minimum	0
Maximum	508
Zeros	1436
Zeros (%)	98.4%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

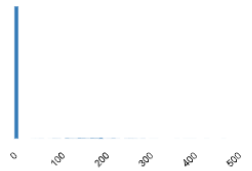
ScreenPorch

Real number (ℝ≥0)

ZEROS

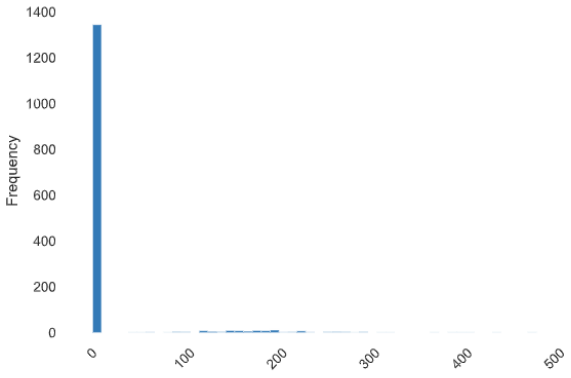
Distinct	76
Distinct (%)	5.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	15.0609589

Minimum	0
Maximum	480
Zeros	1344
Zeros (%)	92.1%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values



Histogram with fixed size bins (bins=50)

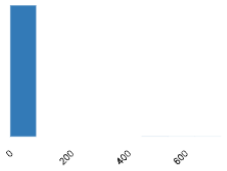
PoolArea

Real number (ℝ≥0)

HIGH CORRELATION
ZEROS

Distinct	8
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.75890411

Minimum	0
Maximum	738
Zeros	1453
Zeros (%)	99.5%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



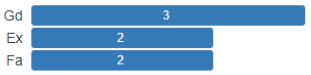
Toggle details

PoolQC

Categorical

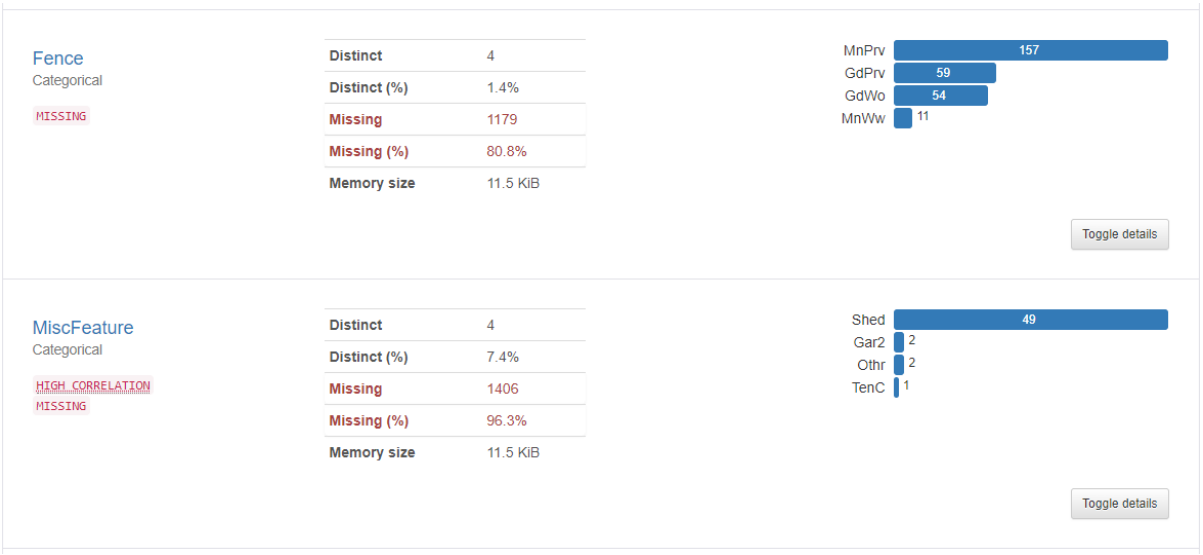
HIGH CORRELATION
MISSING

Distinct	3
Distinct (%)	42.9%
Missing	1453
Missing (%)	99.5%
Memory size	11.5 KiB



Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017



Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

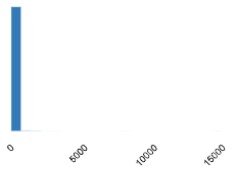
MiscVal

Real number (ℝ≥0)

HIGH CORRELATION
SKEWED
ZEROS

Distinct	21
Distinct (%)	1.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	43.4890411

Minimum	0
Maximum	15500
Zeros	1408
Zeros (%)	96.4%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



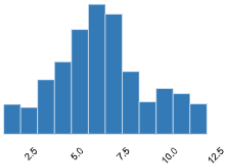
Toggle details

MoSold

Real number (ℝ≥0)

Distinct	12
Distinct (%)	0.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	6.321917808

Minimum	1
Maximum	12
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



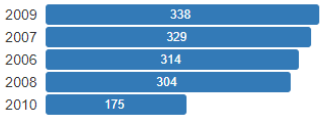
Toggle details

YrSold

Categorical

HIGH CORRELATION

Distinct	5
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



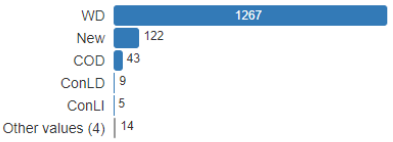
Toggle details

SaleType

Categorical

HIGH CORRELATION

Distinct	9
Distinct (%)	0.6%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB



Toggle details

SaleCondition

Categorical

HIGH CORRELATION

Distinct	6
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	11.5 KiB

Toggle details

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

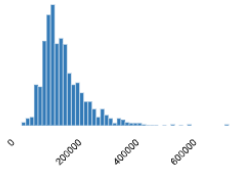
SalePrice

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	663
Distinct (%)	45.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	180921.1959

Minimum	34900
Maximum	755000
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	11.5 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	34900
5-th percentile	88000
Q1	129975
median	163000
Q3	214000
95-th percentile	326100
Maximum	755000
Range	720100
Interquartile range (IQR)	84025

Descriptive statistics

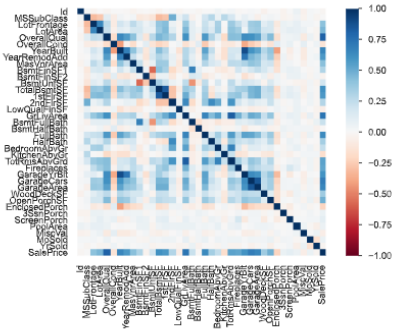
Standard deviation	79442.50288
Coefficient of variation (CV)	0.4391000319
Kurtosis	6.53628186
Mean	180921.1959
Median Absolute Deviation (MAD)	38000
Skewness	1.88287576
Sum	264144946
Variance	6311111264
Monotonicity	Not monotonic

Correlaciones

Correlations

Spearman's ρ | Pearson's r | Kendall's τ | Phik (φk)

Toggle correlation descriptions

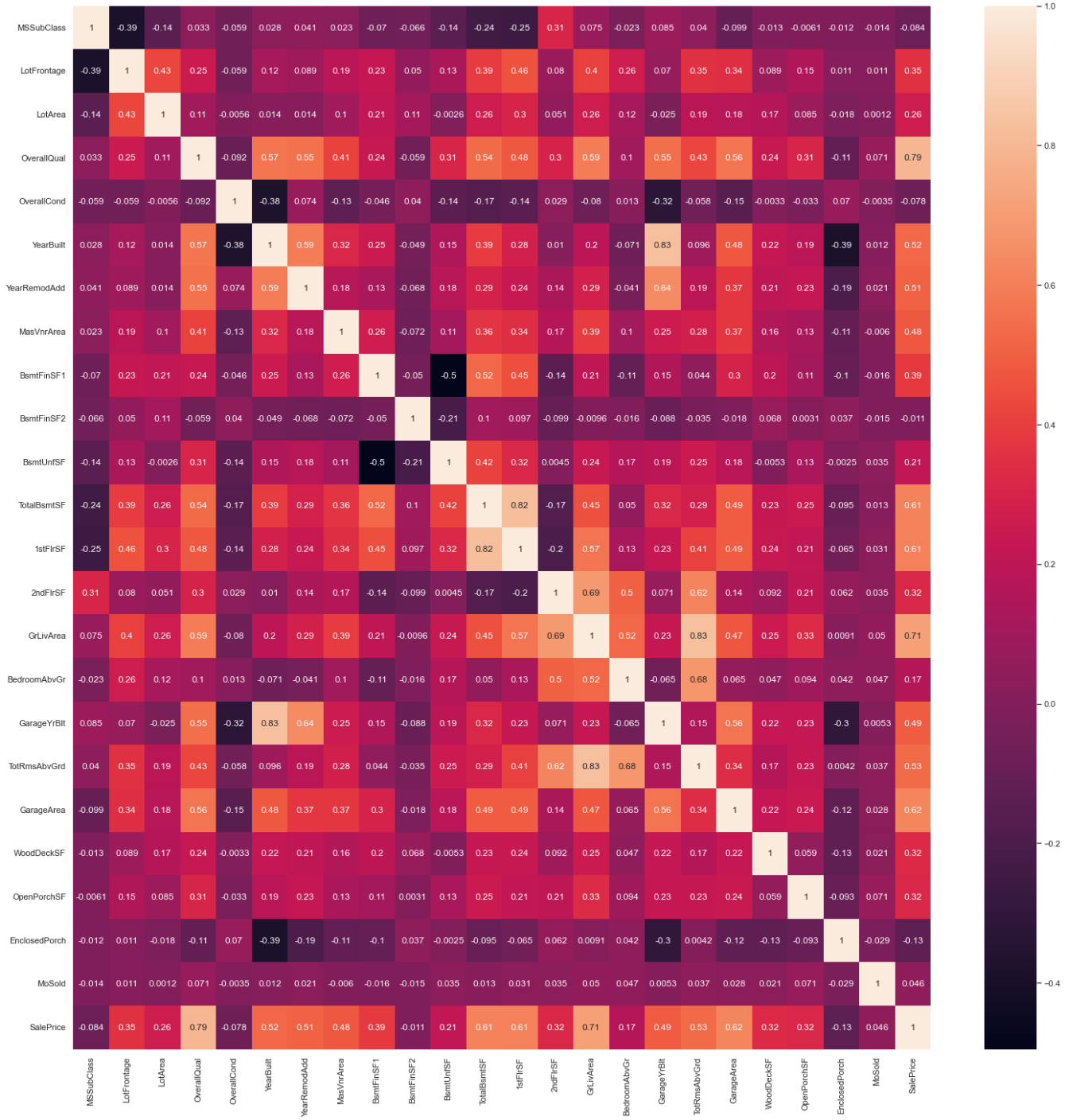


Spearman's ρ

The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's r. It's value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.

To calculate ρ for two variables X and Y, one divides the covariance of the rank variables of X and Y by the product of their standard deviations.

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017



Hallazgos y explicación de procedimiento

Para analizar los datos se utiliza la herramienta de pandas profiler la cual nos brinda información de estadística descriptiva importante de cada una de las variables, al igual que gráficas de barras o histogramas según sea la variable a analizar. Luego de esto se realiza la gráfica de correlación para poder determinar que variables son las que tienen más relación entre ellas y así enfocar más nuestras variables a utilizar.

Analizando los datos podemos observar que muchas de estas variables numéricas están relacionadas con otras. Una de estas por ejemplo puede ser el TotRmsAbvGrd y GrLivArea por que sabemos que dependiendo el área que se tenga va a influir en el total de cuartos

que va a tener. De igual forma logramos determinar que las variables numéricas más importantes para nuestro análisis son: 'MSSubClass', 'LotFrontage', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'BedroomAbvGr', 'GarageYrBlt', 'TotRmsAbvGrd' ,'GarageArea', 'WoodDeckSF' ,'OpenPorchSF', 'EnclosedPorch', 'MoSold', 'SalePrice'. Por qué son algunas de las que tienen relación entre ellas además de que brindan información adecuada. De igual manera logramos determinar que las variables categoricas más importantes son: 'Street','Alley','LotShape','LandContour','LotConfig','LandSlope','BldgType','HouseStyle','BsmtQual','BsmtExposure','HeatingQC','GarageCars'.

Análisis de componentes principales

Análisis factorial

```
    from factor_analyzer.factor_analyzer import calculate_kmo
    kmo_all,kmo_model=calculate_kmo(houses_clean)
    kmo_all, kmo_model
```

```
8]  ✓ 17.4s
```

```
c:\Users\ALIEWARE\AppData\Local\Programs\Python\Python39\lib\site-packages\fa
  warnings.warn('The inverse of the variance-covariance matrix '

(array([0.67986034, 0.85336964, 0.86511003, 0.91950297, 0.52939275,
        0.76092737, 0.83176101, 0.94283582, 0.70914247, 0.11991146,
        0.64250957, 0.87220836, 0.62952928, 0.45033991, 0.66780668,
        0.77086765, 0.82586225, 0.91502783, 0.91616437, 0.94585033,
        0.93825884, 0.71845741, 0.46740089, 0.92301962]),
 0.7683818262377667)
```

Bartlett

```
    from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity

    chi_square_value,p_value=calculate_bartlett_sphericity(houses_clean)
    chi_square_value, p_value
```

```
7]  ✓ 0.9s
```

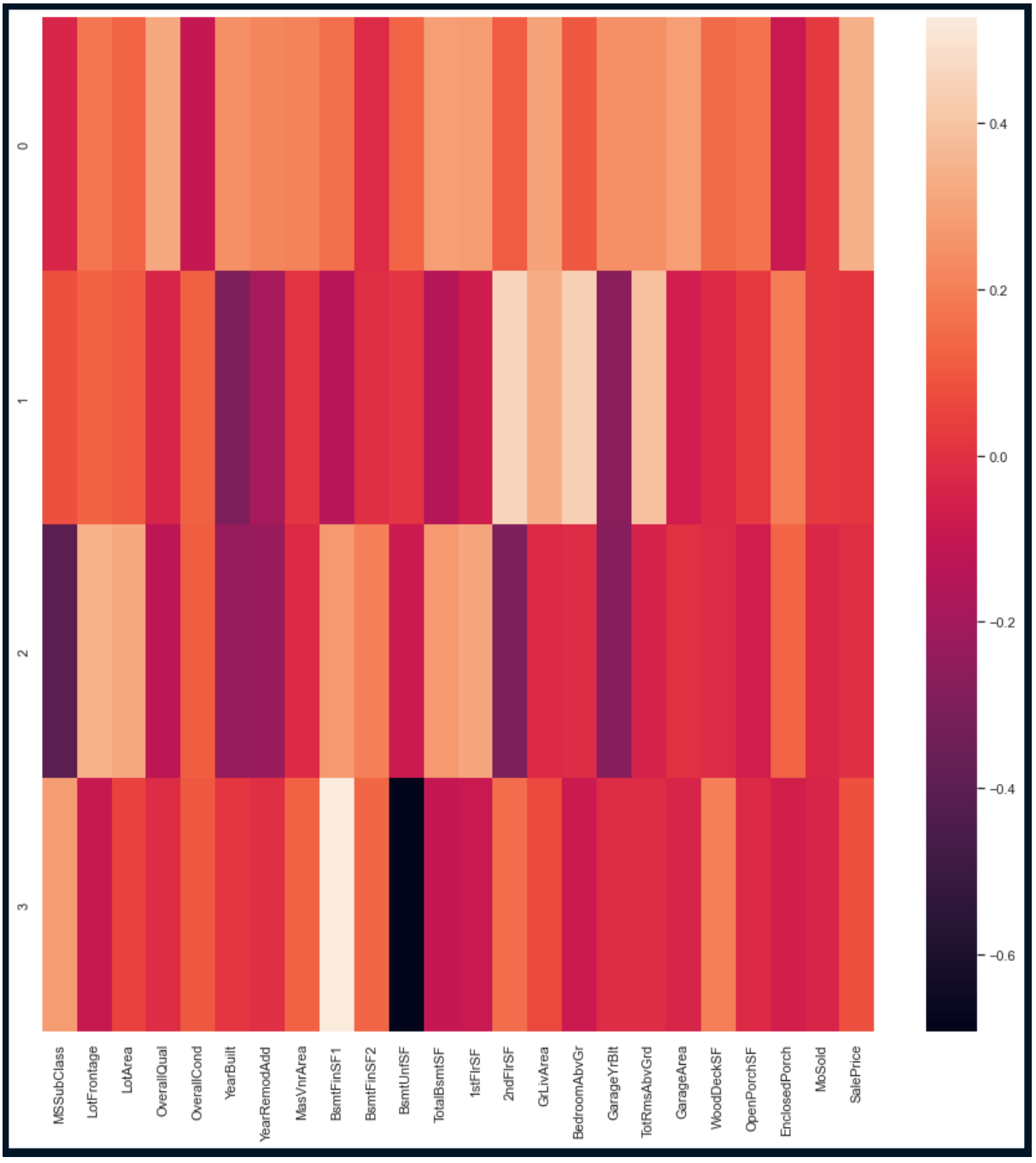
```
(56989.803679788645, 0.0)
```

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Hallazgos

Como se puede observar nuestro resultado de KMO es de 0.77, lo cual nos indica que tenemos datos adecuados para poder seguir con el análisis de PCA.
De igual forma nuestro test de bartlett nos dio 0.0 en el p-value lo cual nos indica que la matriz no es igual a la matriz de identidad por lo cual se puede continuar con el análisis, de igual forma nos indica que el valor de chi-cuadrado es de 56989.80

Matriz de correlación



Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

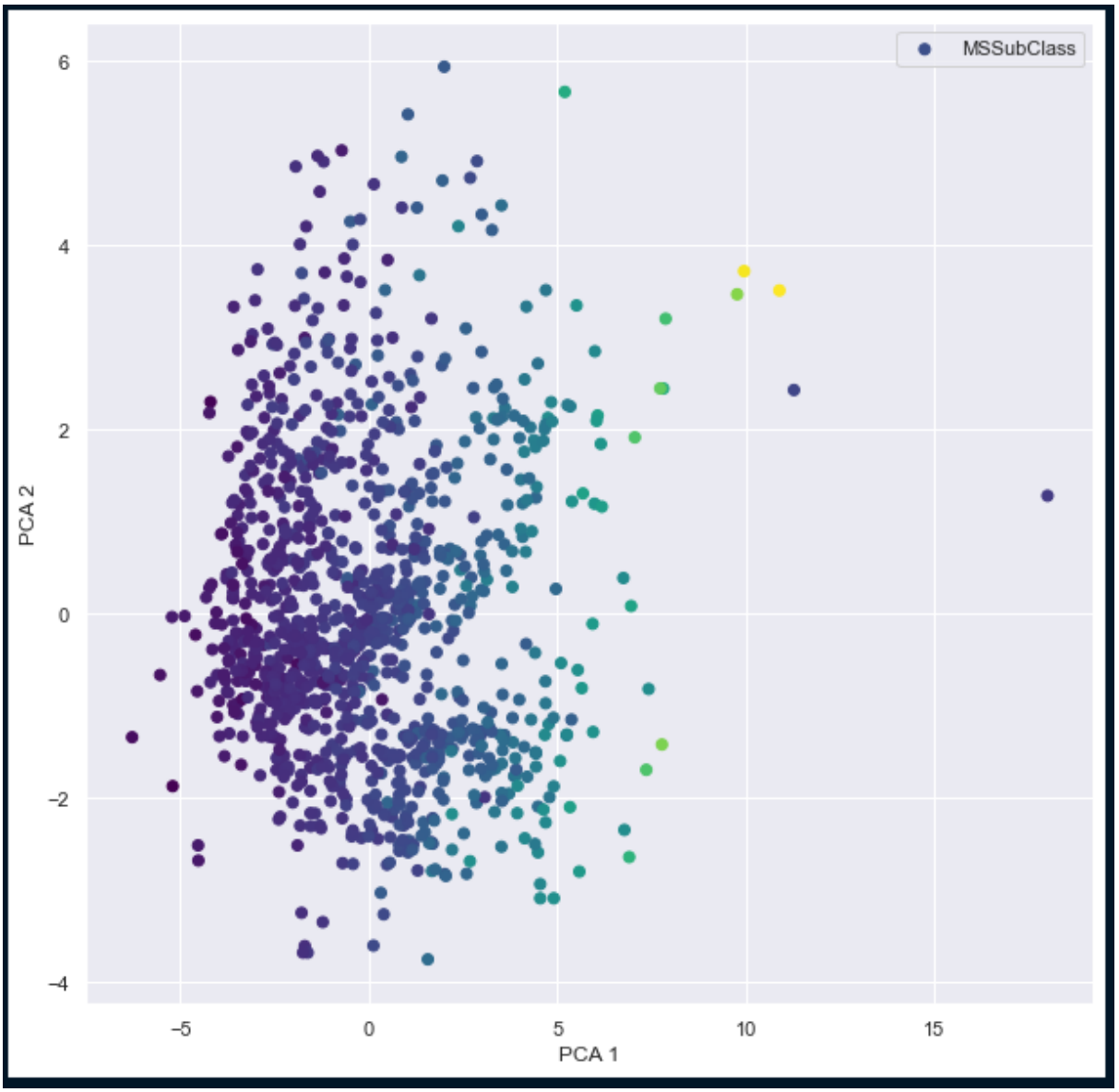
Como podemos observar nuestra matriz tiene 4 filas, eso se decide de esta forma debido a que en nuestra gráfica de PCA podemos observar alrededor de 4 distintos colores. Para lograr determinar cuales son las variables que tienen más relación entre sí primero lo que se hizo fue separar las cualitativas más importantes y luego con ayuda del algoritmo de PCA determinar cuántos componente iba a tener. En nuestro caso los 4 principales componentes que se observaron.

Coeficientes principales.

Analizando nuestros componentes principales logramos identificar que el primero (fila 0) tiene una fuerte relación con GarageArea,GrlivArea, 1flrSF, TotalBsmtSF, OveralQuall entre algunas otras pero principalmente con estas 5 variables.
Nuestro segundo componente tiene una gran relación con 2ndFlrSF, BedroomAbvGrnd, TotRmsAbvGrnd y GrlivArea.
En nuestro caso el tercero se puede observar 5 componentes LotFrontArea, LotArea, BsmrFinF1,TotalBsmtSf y 1stFlrSF.
Por último nuestro tercer componente se puede observar que está fuertemente relacionado con BsmFinSF1 seguido de MSSubClass y WoodDeckSF.

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

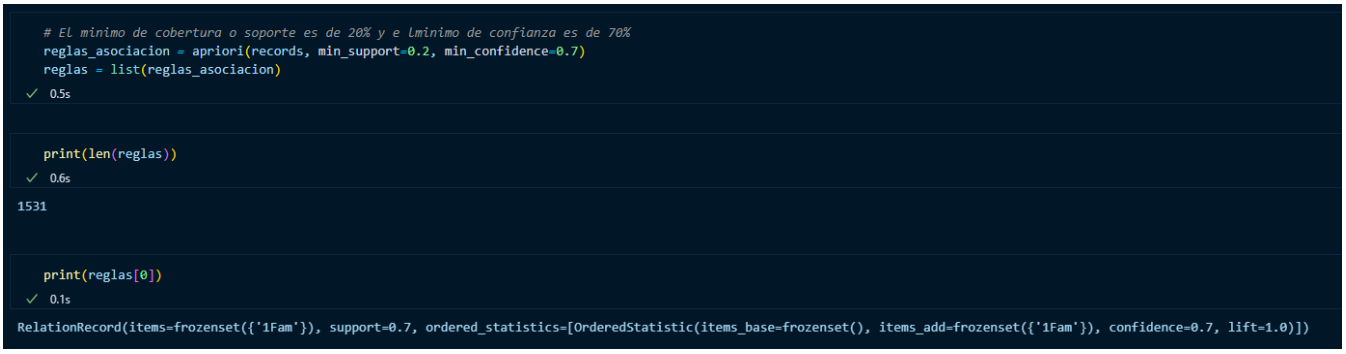
PCA



Observando nuestra gráfica de PCA podemos determinar que existen entre 4 y 5 componentes principales. Esto quiere decir que entre todas las observaciones que tenemos existen 4 principales. En nuestro caso vamos a tomar 4 como principales debido a que se pueden observar 4 colores principalmente destacados en la gráfica.

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

Reglas de asociación



Analizando las reglas de asociación nos dimos cuenta que de todos los datos (118260) se pueden crear 1531 reglas. Esto se puede decir que existen alrededor de 1531 distintas casas y todas las casas dentro de nuestra base de datos entra en alguna de estas reglas de asociación. En resumen se crean 1,531 reglas con las variables categóricas proporcionadas. Lo cual lo podemos interpretar como 1,531 distintos tipos de casas. Con esto logramos encontrar el conjunto de datos frecuentes.

Hallazgos y conclusiones

Para resumir el proceso podemos decir que lo principal que se debe de realizar es el análisis de las variables. Una vez clasificadas las variables será más fácil poder determinar qué variables tienen más relación entre sí, qué variables son más relevantes. En este caso se utilizó la herramienta de pandas profile que nos permite realizar un análisis más detallado de todos estos datos. Cantidades mínimas, máximas, al igual que cuartiles y varianzas, entre otras.
Gracias a este análisis de variables logramos utilizar el algoritmo PCA. Este nos ayudó a determinar que existen cuatro componentes principales en nuestro conjuntos de datos y con ayuda de la matriz de correlación logramos determinar cuales son las variables cuantitativas que tienen más relación con dichos componentes principales.
Por último con ayuda del algoritmo a priori logramos formar un conjunto de reglas de asociación. Estas nos dicen que hay 1,531 distintos sets de casas con las mismas características. Estos datos nos ayudan a encontrar el conjunto de datos frecuentes en nuestra base de datos, en este caso las casas con propiedades similares.

Rúbrica

(42 puntos) Análisis Exploratorio:

- ☑ ~~Estudia las variables cuantitativas mediante técnicas de estadística descriptiva~~
- ☑ ~~Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión que ayudan a explicar los datos~~
- ☑ ~~Analiza las correlaciones entre las variables, trata de explicar los outliers (puntos atípicos) y toma decisiones acertadas ante la presencia de valores faltantes.~~
- ☑ ~~Estudia las variables categóricas~~
- ☑ ~~Elabora gráficos de barra, tablas de frecuencia y de proporciones~~
- ☑ ~~Elabora gráficos adecuados según el tipo de dato que representan~~

Donaldo Sebastian Garcia Jiménez 19683
Raul Angel Jimenez Hernandez 19017

- ☑ ~~Explica muy bien todos los procedimientos y los hallazgos que va haciendo.~~

(18 puntos) Análisis de componentes Principales

- ☑ ~~Estudia la matriz de correlación, la agrega y explica lo que observa en ella~~
- ☑ ~~Determina si es posible usar la técnica de análisis factorial para hallar las componentes principales~~
- ☑ ~~Determina si vale la pena aplicar las componentes principales interpretando el test de esfericidad de Bartlett~~
- ☑ ~~Obtiene los componentes principales y explica cuántos seleccionará para explicar la mayor variabilidad posible.~~
- ☑ ~~Interpreta los coeficientes principales.~~

(18 puntos) Reglas de asociación

- ☑ ~~Construye reglas de asociación usando el algoritmo a priori.~~
- ☑ ~~Discute sobre las reglas de asociación más interesantes teniendo en cuenta sus niveles de confianza y soporte.~~

(22 puntos) Hallazgos y conclusiones.

- ☑ ~~Hace un resumen de los hallazgos en el análisis exploratorio~~
- ☑ ~~Llega a conclusiones sobre el análisis de componentes principales~~
- ☑ ~~Determina las reglas de asociación más interesantes.~~