Laboratorio 1.

Data Science

Maria José Morales 19145

Luis Pedro García 19344

**1. Haga una exploración rápida de sus datos para eso haga un resumen de su dataset.**

Se hizo una breve exploración con la función “dim ()” y la función“summary ()” para ver las dimensiones del data set y un breve resumen de las variables.

En el data set train hay 47 variables cualitativas y 34 variables cuantitativas. Se encontraron algunas irregularidades en los nombres de las variables ya que algunas veces se pone “AbvGrd” para referirse a “Above Grade” y a veces se pone “AvdGr”.

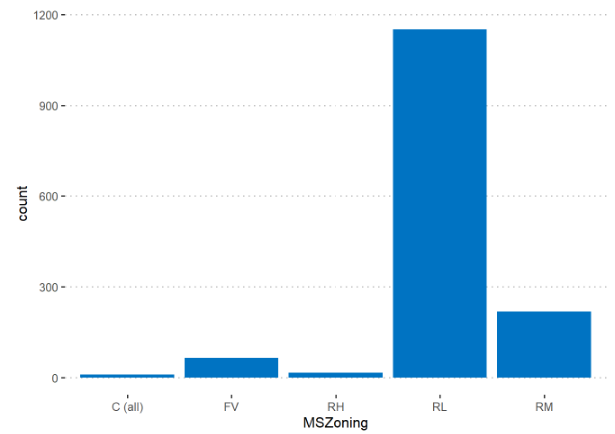
**2. Diga el tipo de cada una de las variables del dataset (cualitativa o categórica, cuantitativa continua, cuantitativa discreta)**

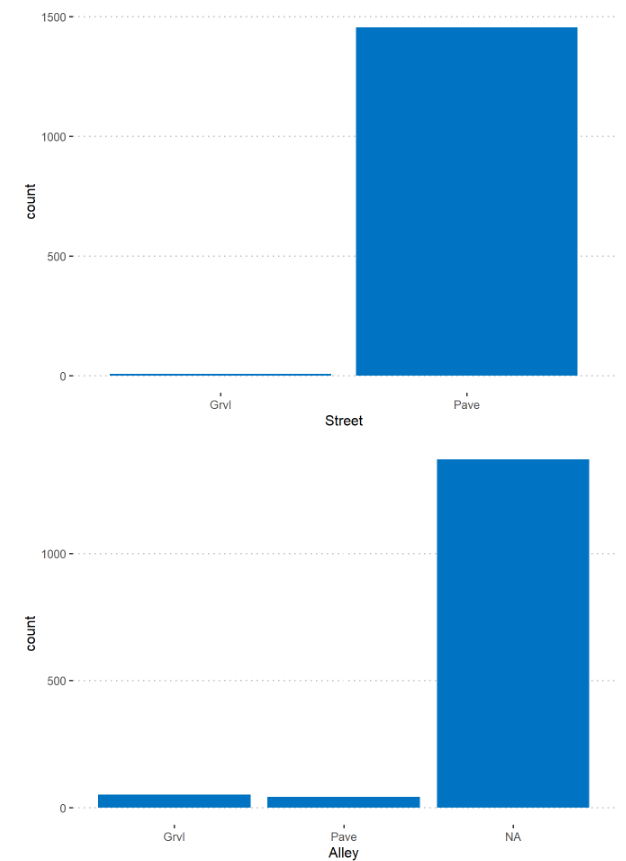
Las variables junto con su categoría son las siguientes:

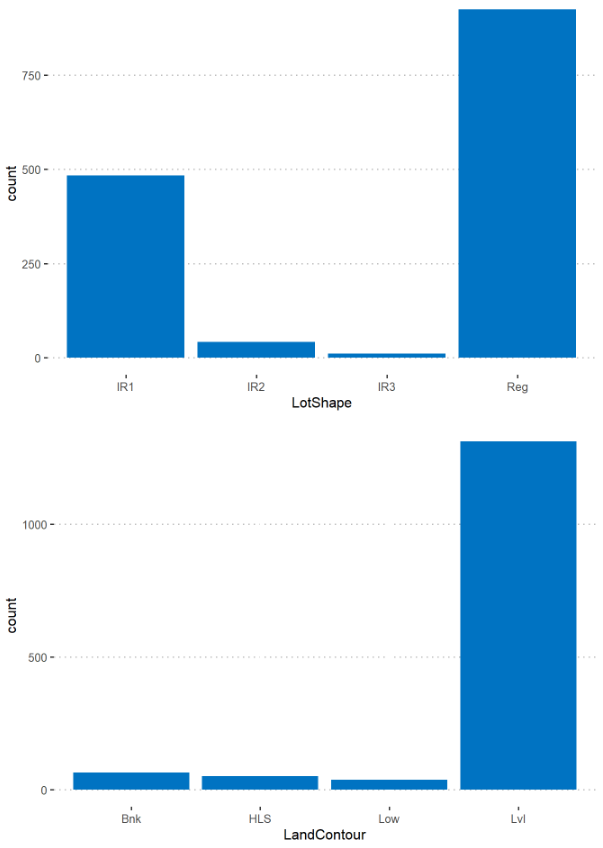
1. Id (cuantitativa discreta)
2. MSSubClass (cuantitativa discreta)
3. MSZoning (cualitativa)
4. LotFrontage(cuantitativa continua)
5. LotArea(cuantitativa continua)
6. Street(cualitativa)
7. Alley(cualitativa)
8. LotShape(cualitativa)
9. LandContour(cualitativa)
10. Utilities(cualitativa)
11. LotConfig(cualitativa)
12. LandSlope(cualitativa)
13. Neighborhood(cualitativa)
14. Condition1(cualitativa)
15. Condition2(cualitativa)
16. BldgType(cualitativa)
17. HouseStyle(cualitativa)
18. OverallQual (cuantitativa continua)
19. OverallCond(cuantitativa continua)
20. YearBuilt(cualitativa)
21. YearRemodAdd(cualitativa)
22. RoofStyle(cualitativa)
23. RoofMatl(cualitativa)
24. Exterior1st(cualitativa)
25. Exterior2nd(cualitativa)
26. MasVnrType(cualitativa)
27. MasVnrArea(cuantitativa continua)
28. ExterQual(cualitativa)
29. ExterCond(cualitativa)
30. Foundation(cualitativa)
31. BsmtQual(cualitativa)
32. BsmtCond(cualitativa)
33. BsmtExposure(cualitativa)
34. BsmtFinType1(cualitativa)
35. BsmtFinSF1(cuantitativa continua)
36. BsmtFinType2(cualitativa)
37. BsmtFinSF2(cuantitativa continua)
38. BsmtUnfSF(cuantitativa continua)
39. TotalBsmtSF(cuantitativa continua)
40. Heating(cualitativa)
41. HeatingQC(cualitativa)
42. CentralAir(cualitativa)
43. Electrical(cualitativa)
44. 1stFlrSF(cuantitativa continua)
45. 2ndFlrSF(cuantitativa continua)
46. LowQualFinSF(cuantitativa continua)
47. GrLivArea(cuantitativa continua)
48. BsmtFullBath(cuantitativa discreta)
49. BsmtHalfBath(cuantitativa discreta)
50. FullBath(cuantitativa discreta)
51. HalfBath(cuantitativa discreta)
52. BedroomAbvGr(cuantitativa discreta)
53. KitchenAbvGr(cuantitativa discreta)
54. KitchenQual(cualitativa)
55. TotRmsAbvGrd(cuantitativa discreta)
56. Functional(cualitativa)
57. Fireplaces(cuantitativa discreta)
58. FireplaceQu(cualitativa)
59. GarageType(cualitativa)
60. GarageYrBlt(cualitativa)
61. GarageFinish(cualitativa)
62. GarageCars(cuantitativa continua)
63. GarageArea(cuantitativa continua)
64. GarageQual(cualitativa)
65. GarageCond(cualitativa)
66. PavedDrive(cualitativa)
67. WoodDeckSF(cuantitativa continua)
68. OpenPorchSF(cuantitativa continua)
69. EnclosedPorch(cuantitativa continua)
70. 3SsnPorch(cuantitativa continua)
71. ScreenPorch(cuantitativa continua)
72. PoolArea(cuantitativa continua)
73. PoolQC(cualitativa)
74. Fence(cualitativa)
75. MiscFeature(cualitativa)
76. MiscVal(cuantitativa continua)
77. MoSold(cuantitativa discreta)
78. YrSold(cualitativa)
79. SaleType(cualitativa)
80. SaleCondition(cualitativa)
81. SalePrice(cuantitativa continua)

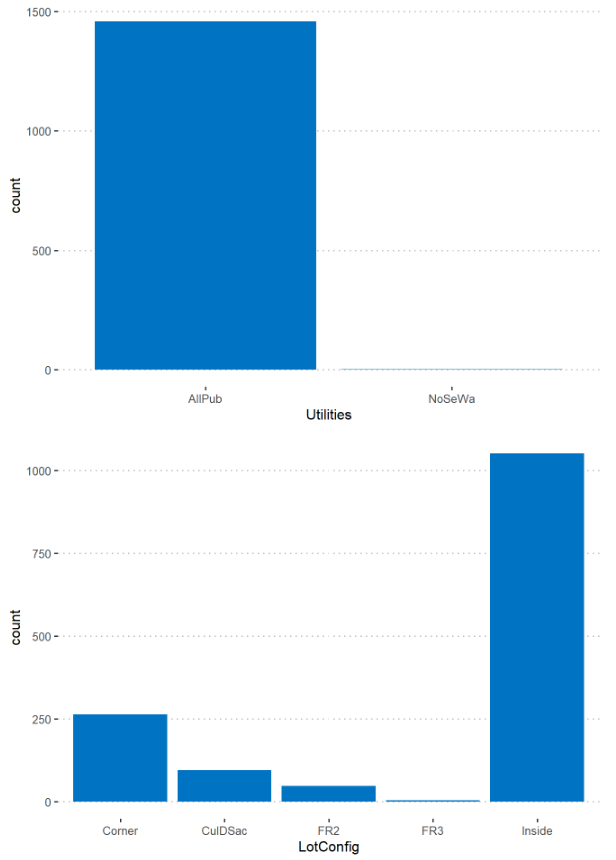
**3. Incluya los gráficos exploratorios siendo consecuentes con el tipo de variable que están representando**

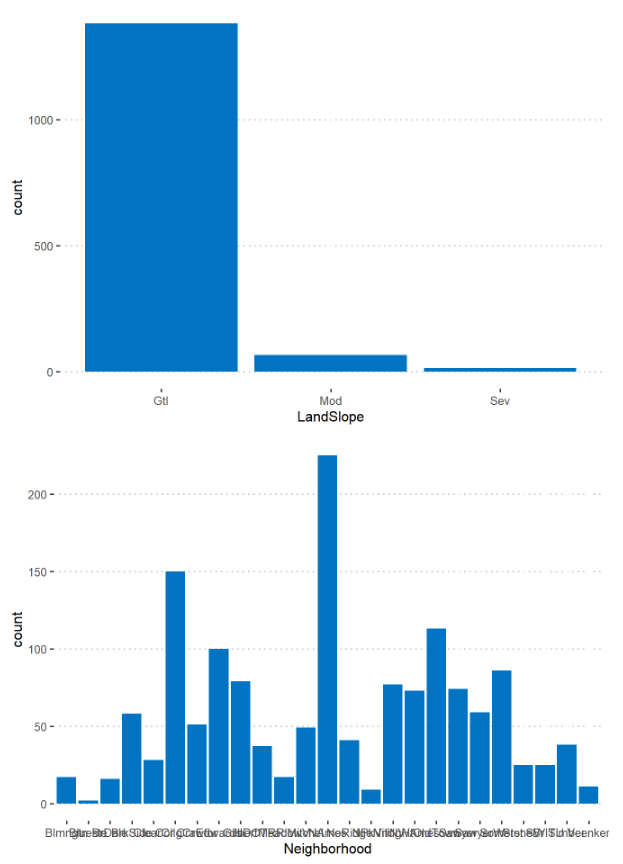
Para las variables cualitativas se hicieron las siguientes tablas de frecuencia:

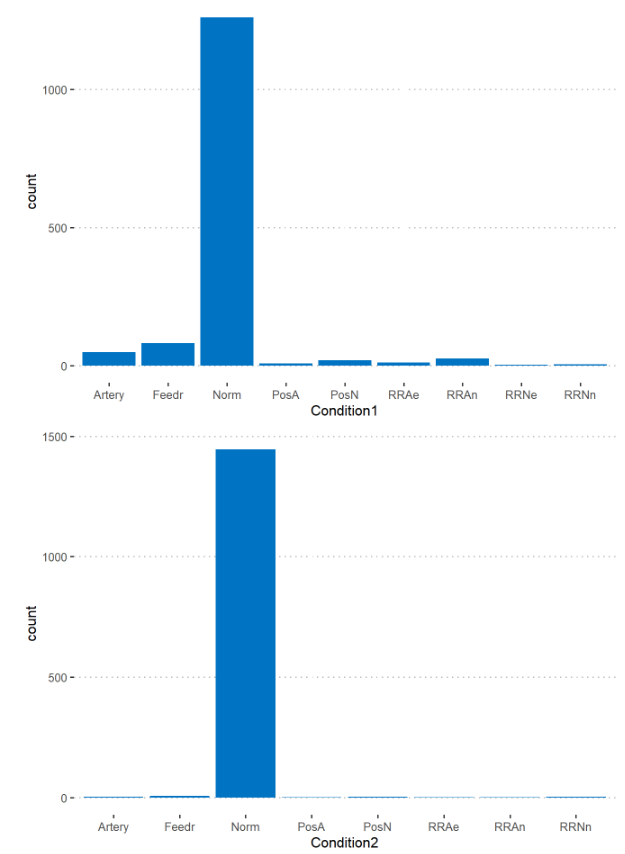


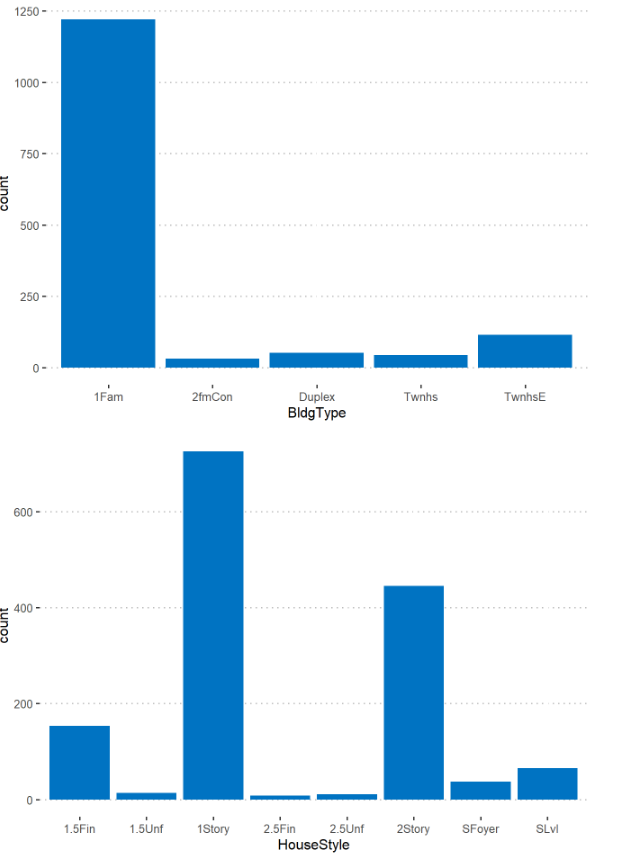


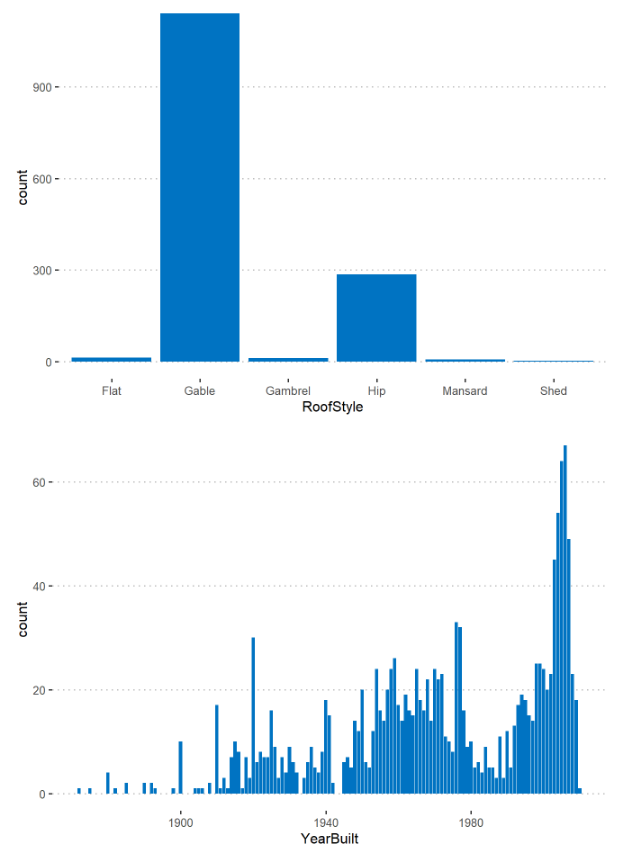




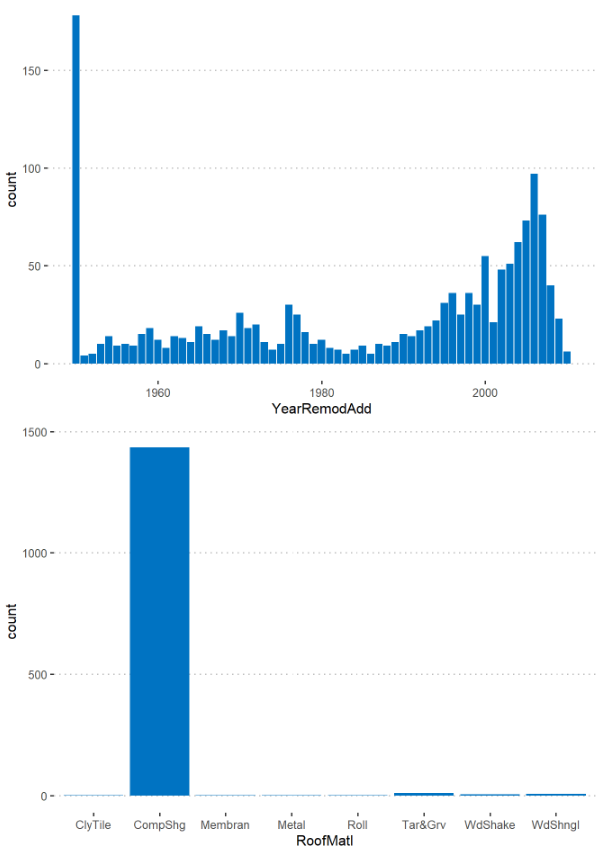


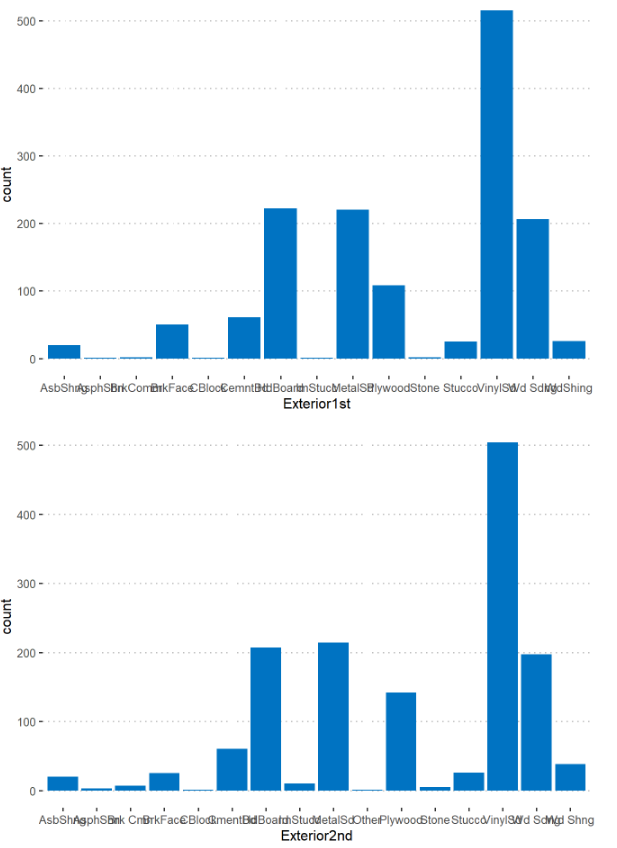


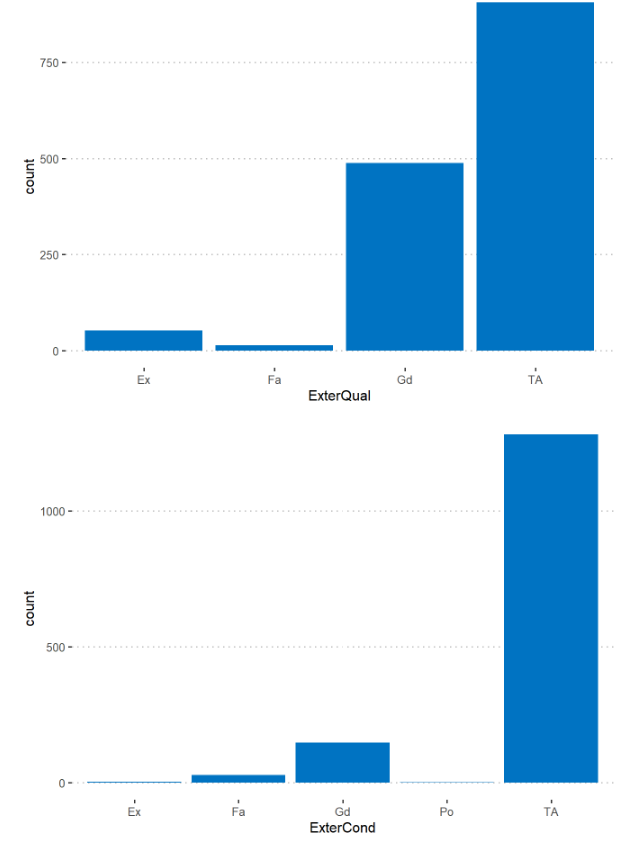


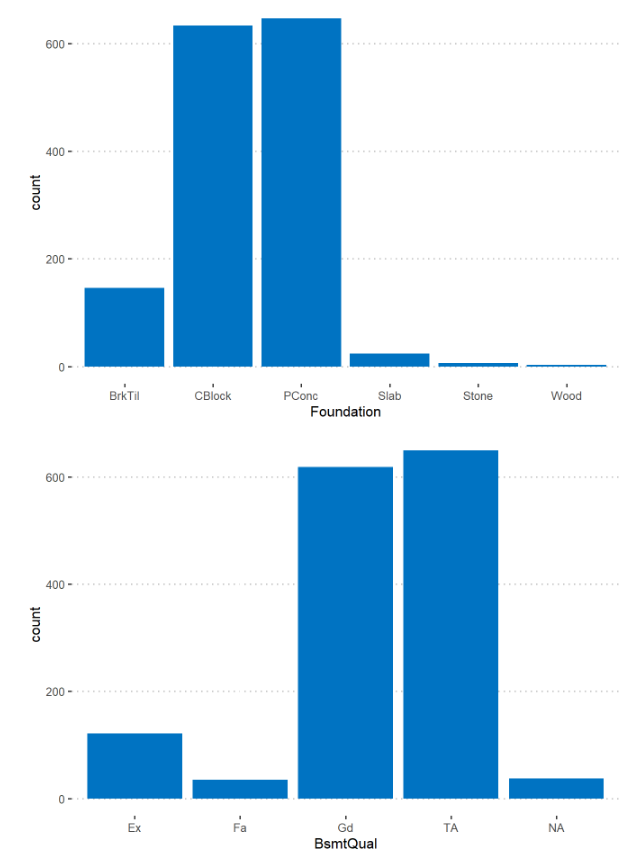


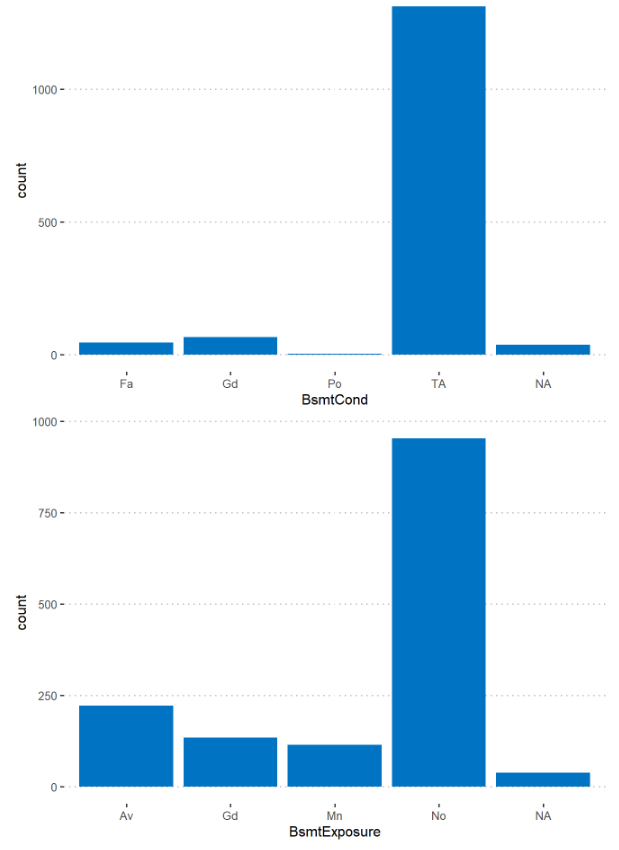
En el año de construcción (YearBuilt) se observa un claro crecimiento en los años recientes, probablemente por el crecimiento poblacional.

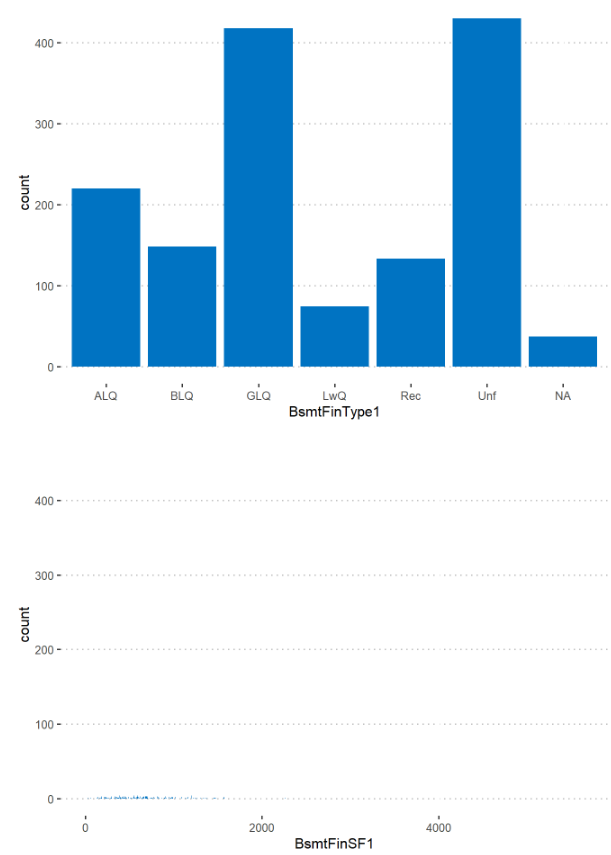


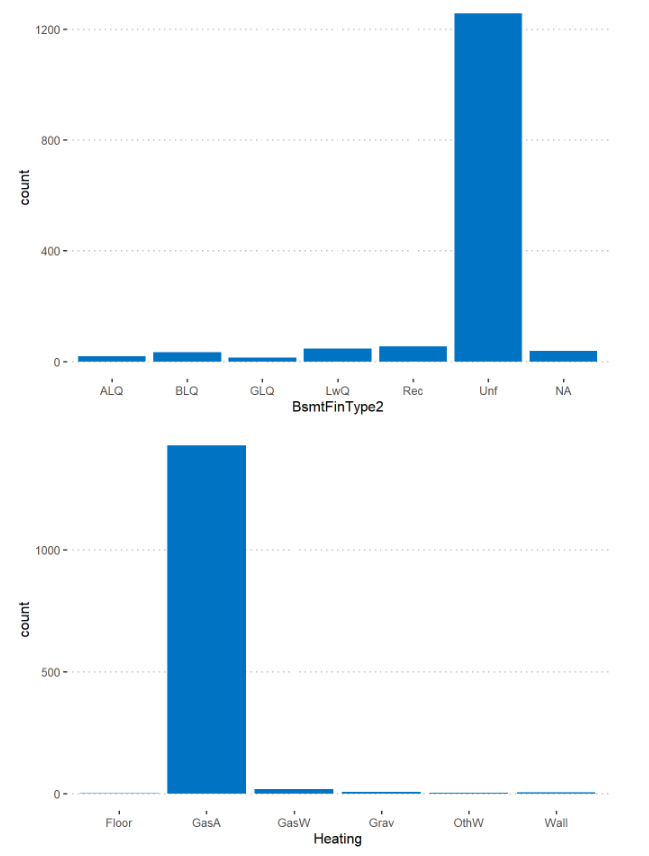


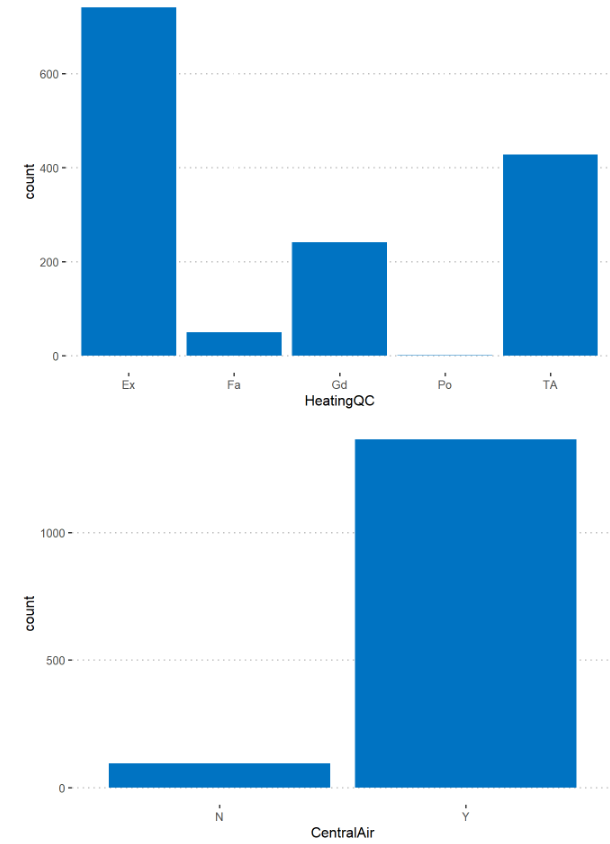


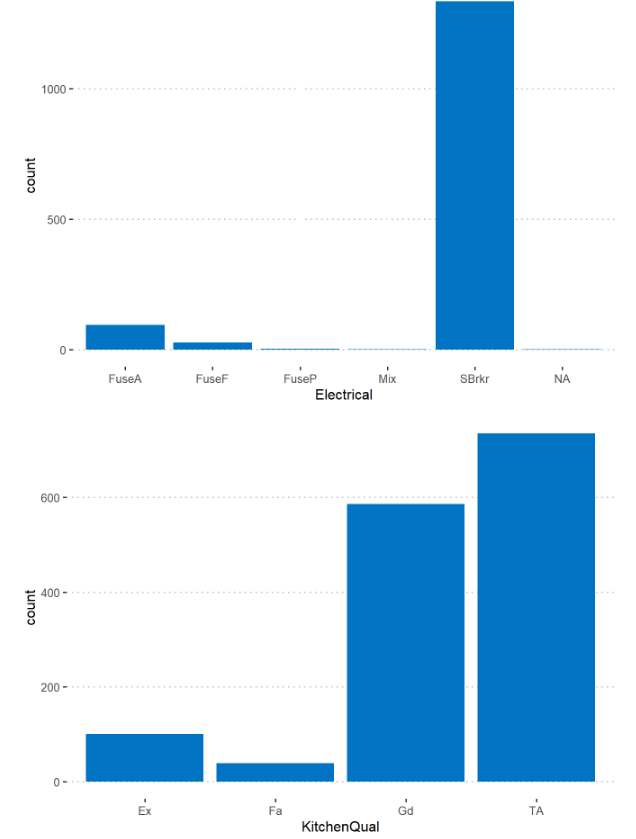


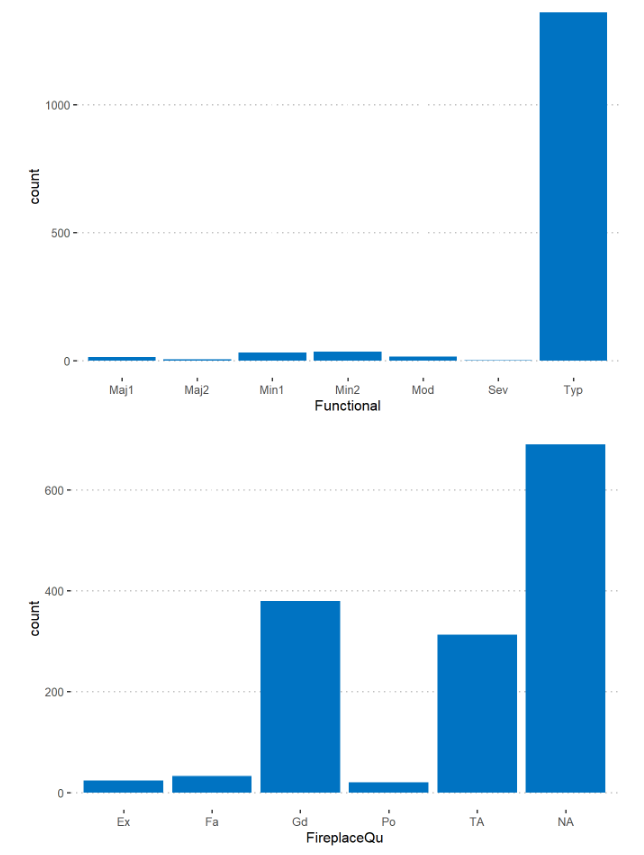


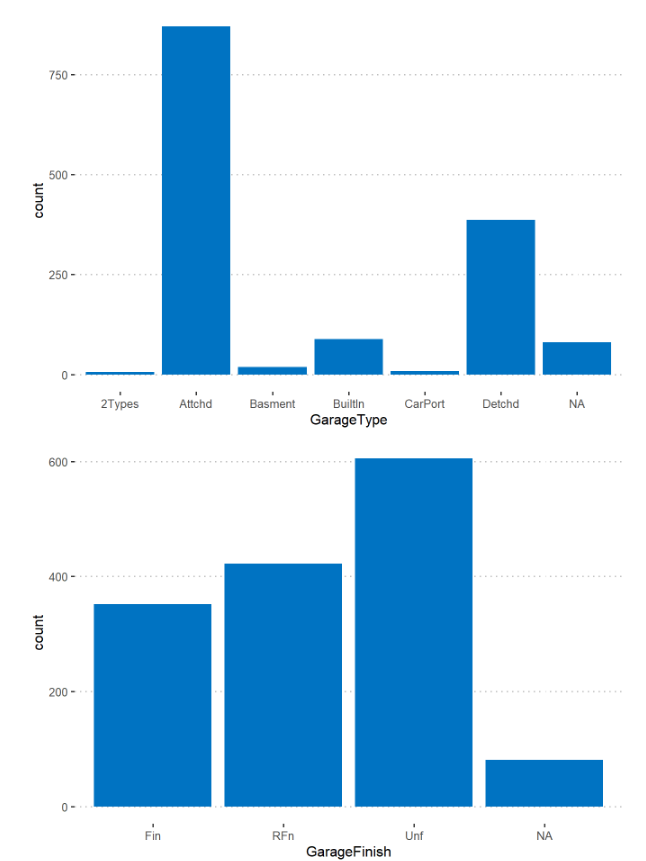


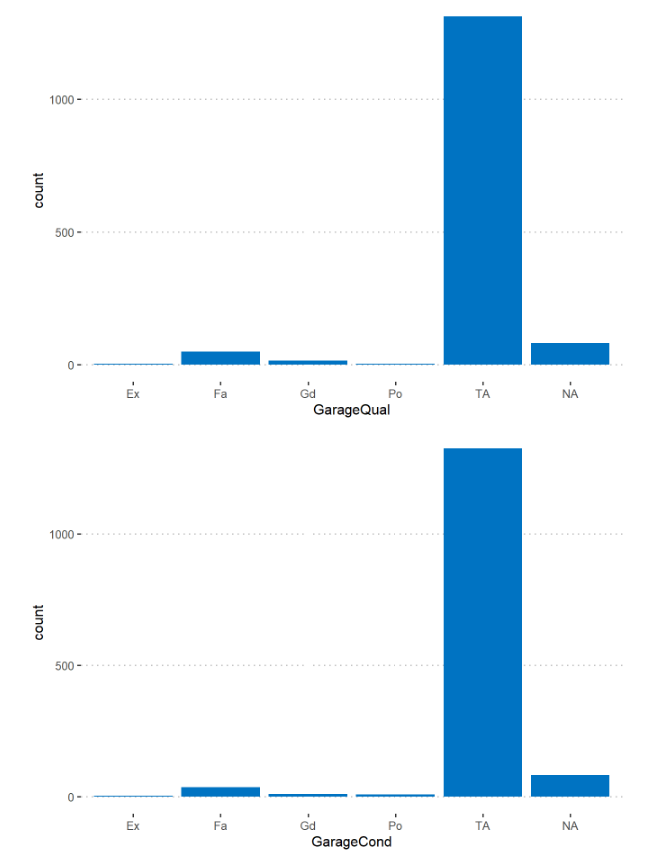


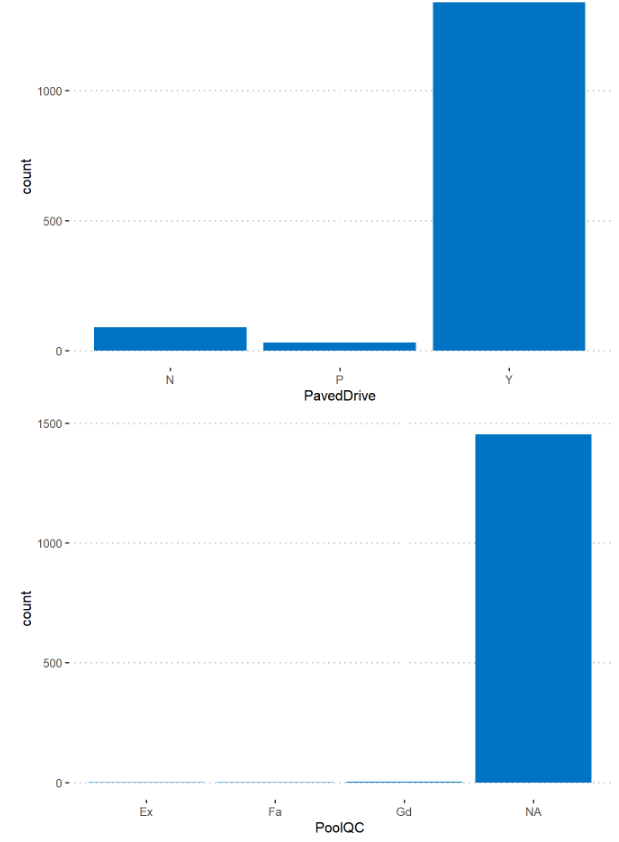


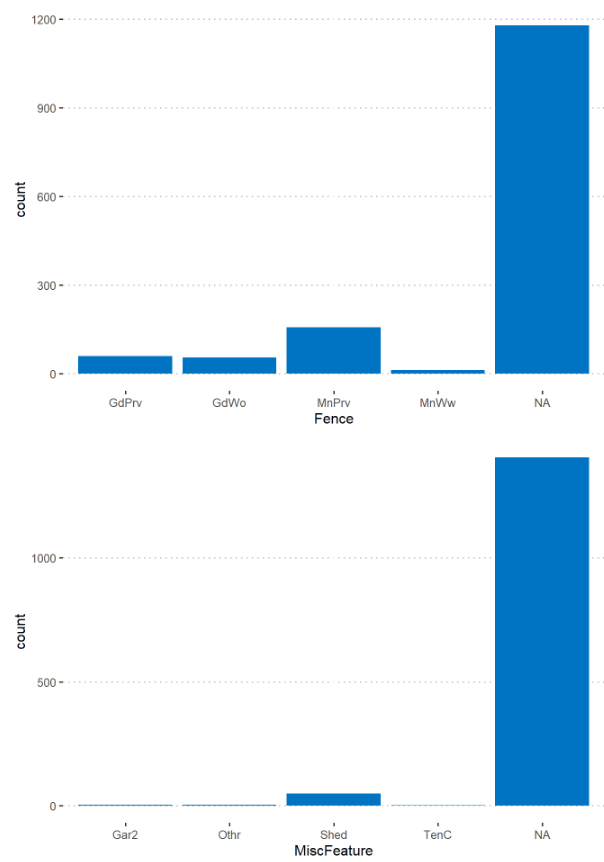


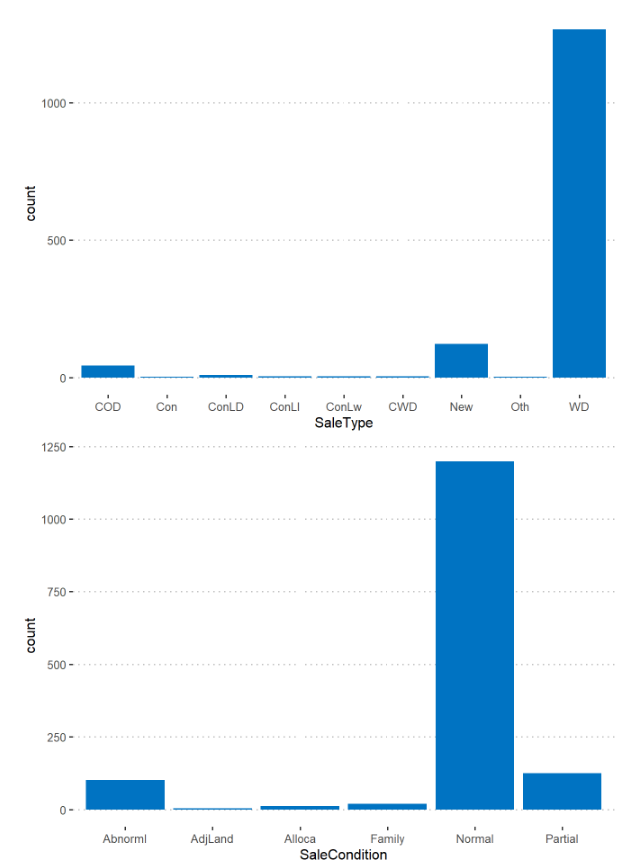




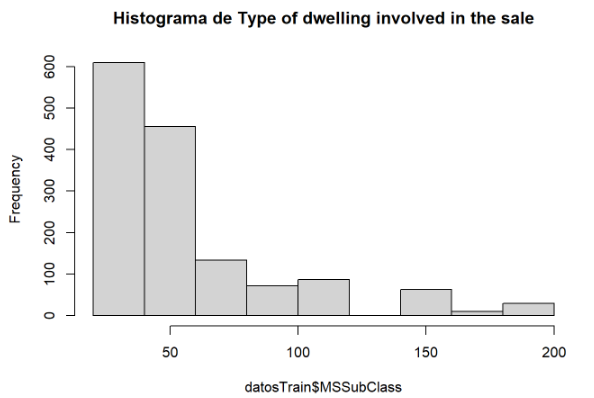


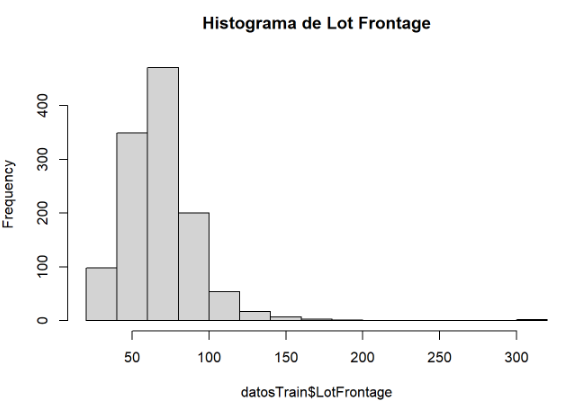




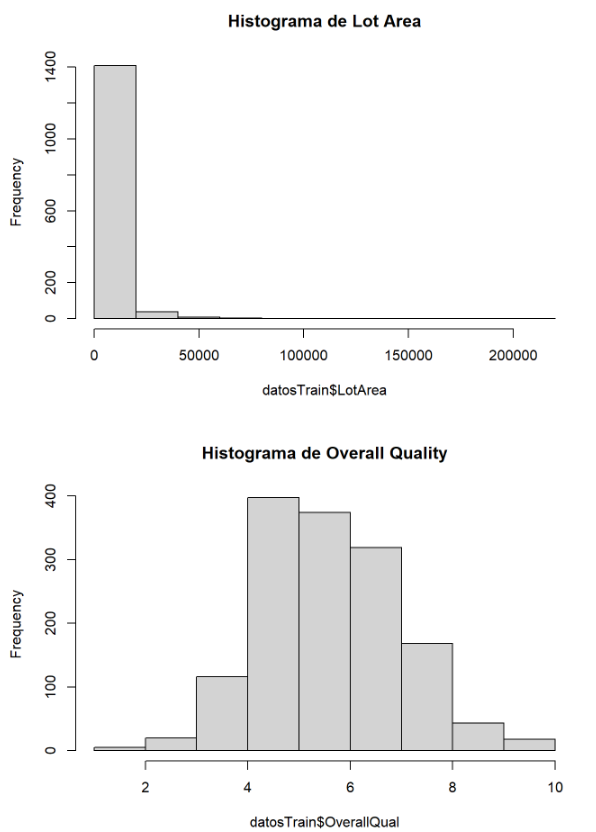


Por otro lado, para las variables cuantitativas, tanto continuas como discretas, se hicieron los siguientes histogramas:

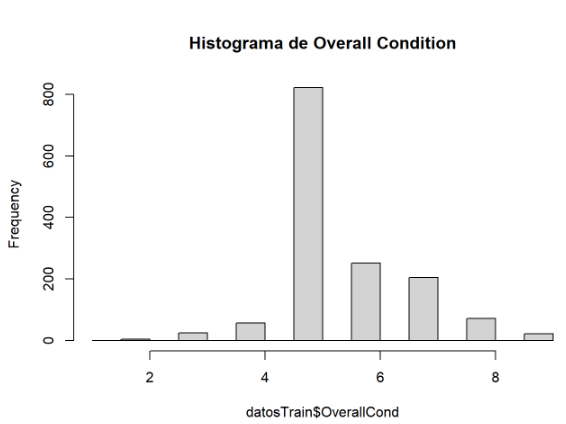




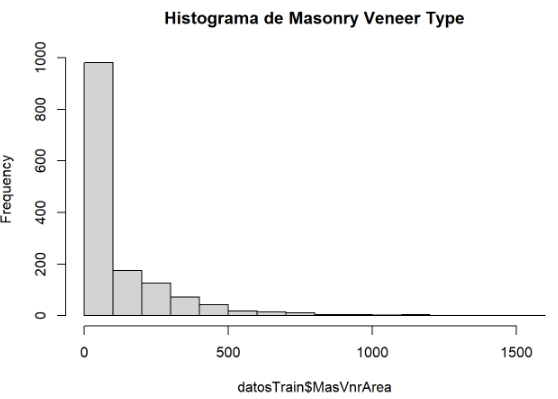
Se puede observar que LotFrontage sigue una distribución normal por su naturaleza continua, que es la distancia del hogar a la calle.

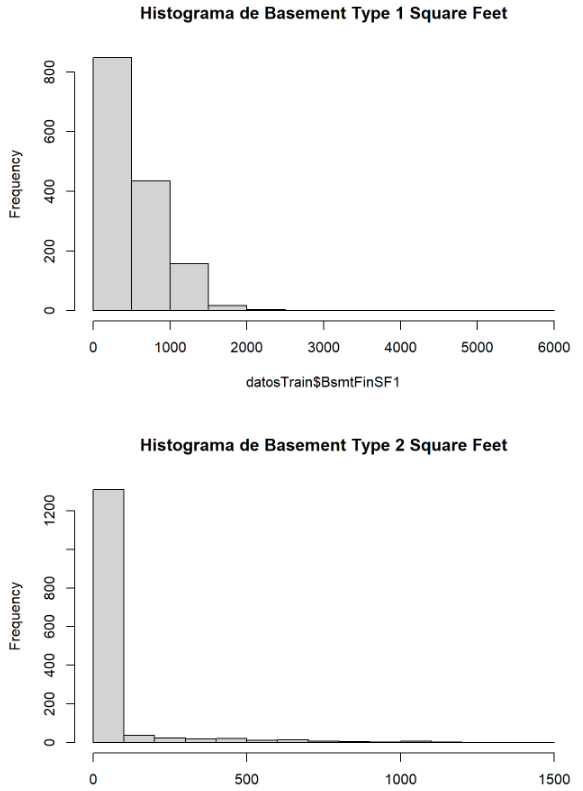


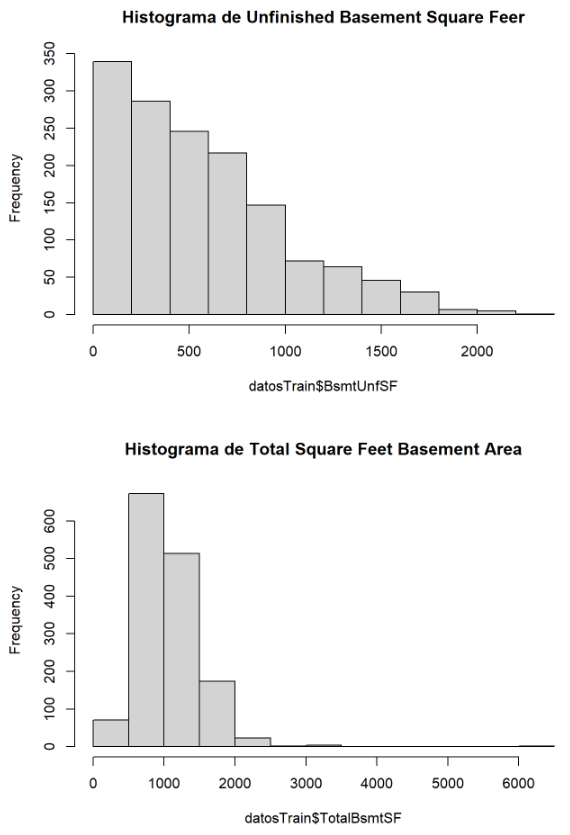
Se observa que Overall Quality también sigue una distribución normal.



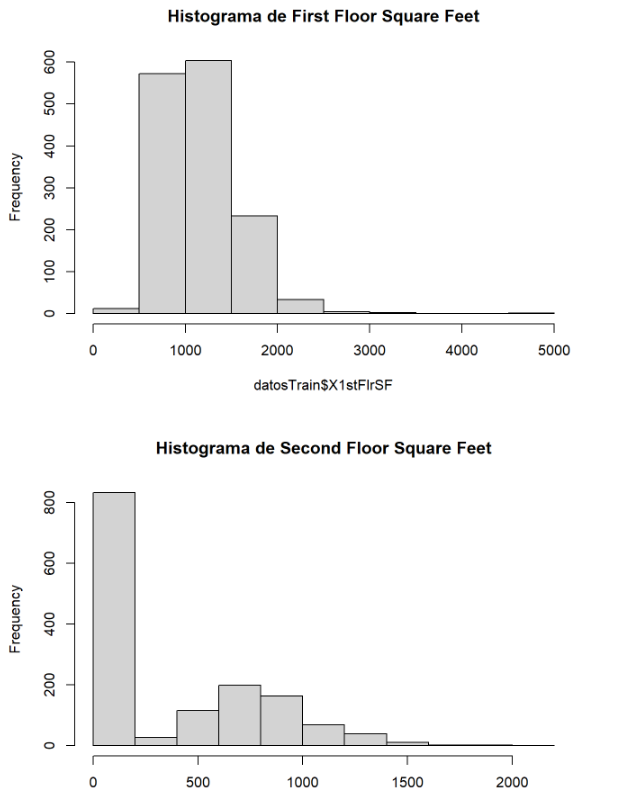
Overall Condition también sigue una distribución normal



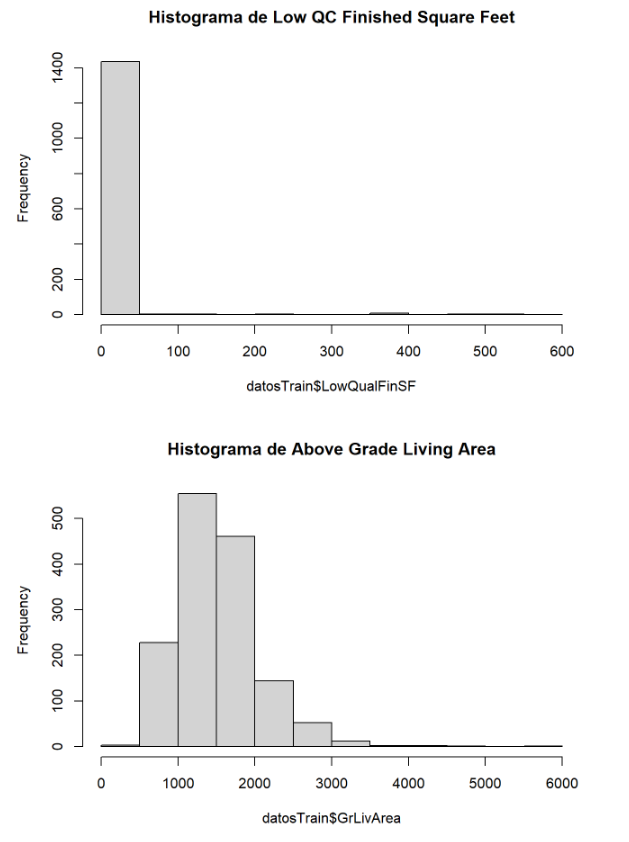




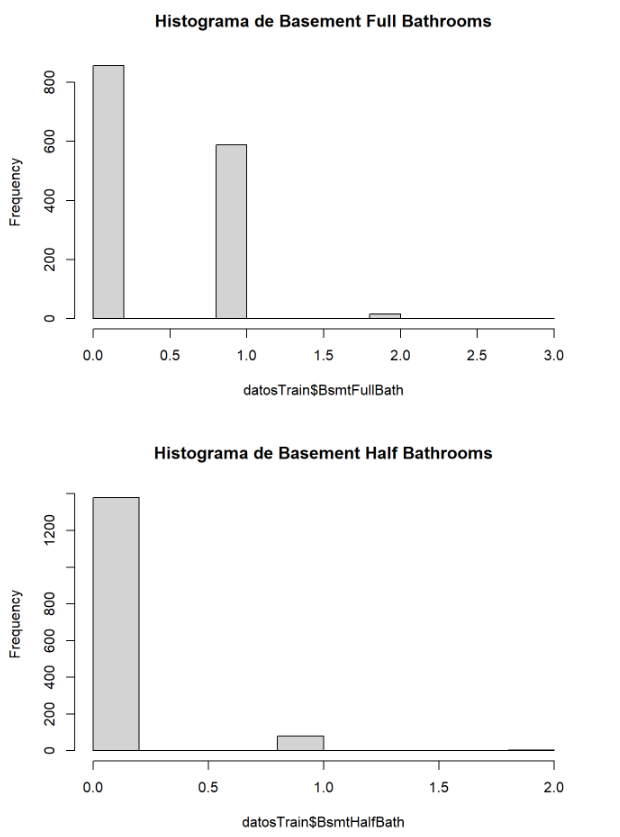
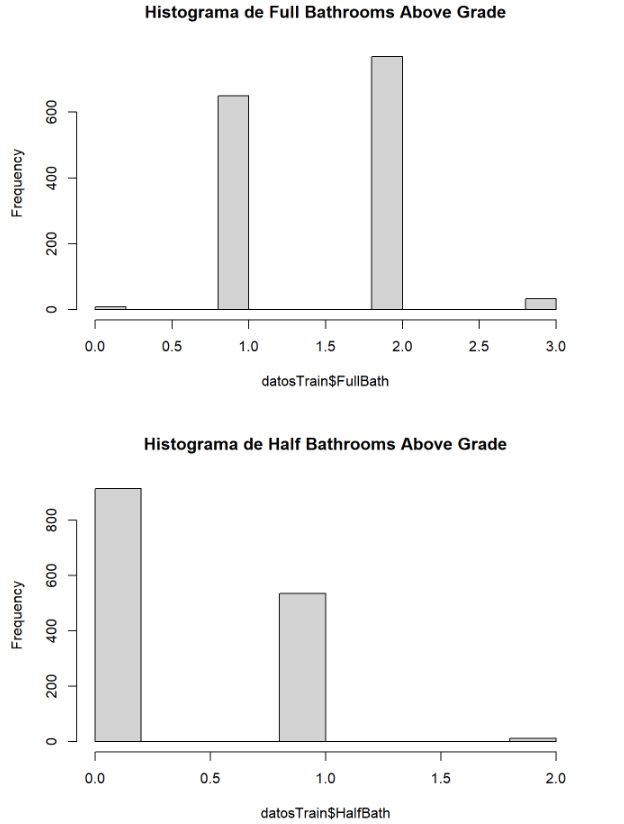
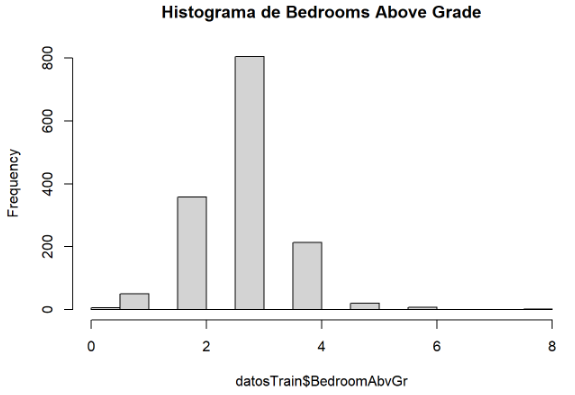
Se observa que Total Square Feet Basement Area también sigue una distribución normal.



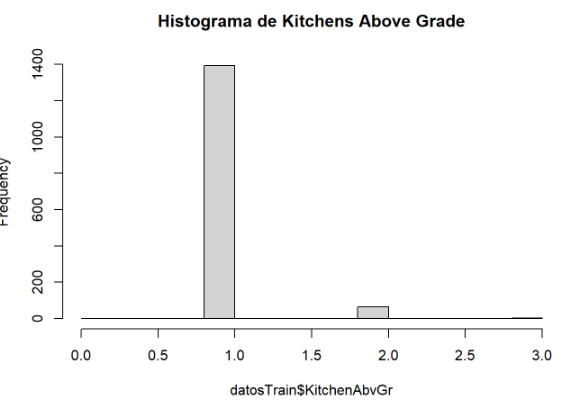
Se observa que First Floor Square Feet sigue una distribución normal.

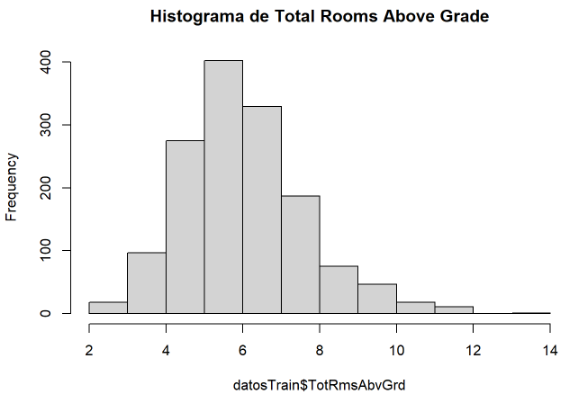


Se observa que Above Grade Living Area tiene una distribución normal.

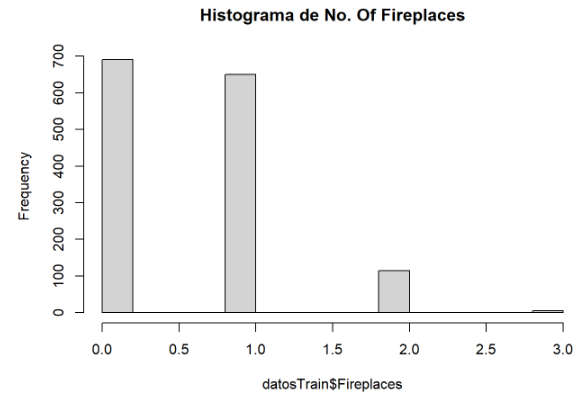
  

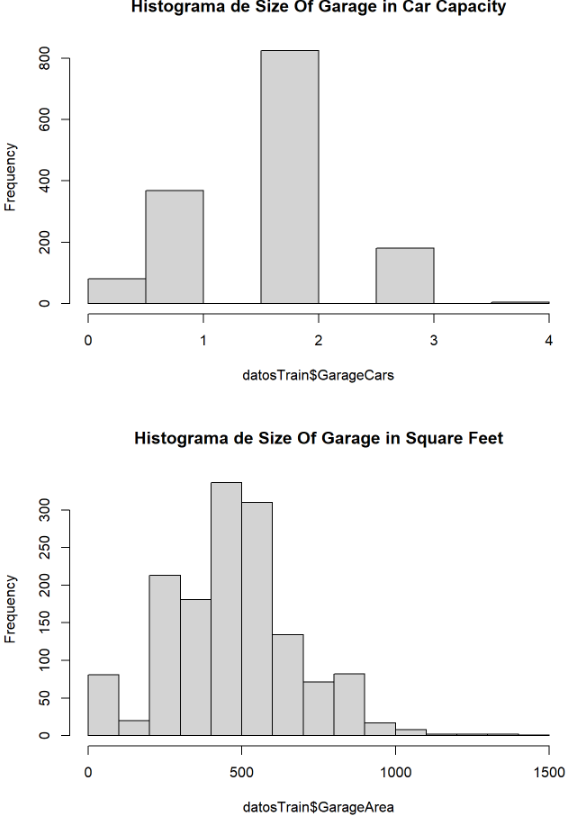
Se observa que Bedrooms Above Grade sigue una distribución normal.



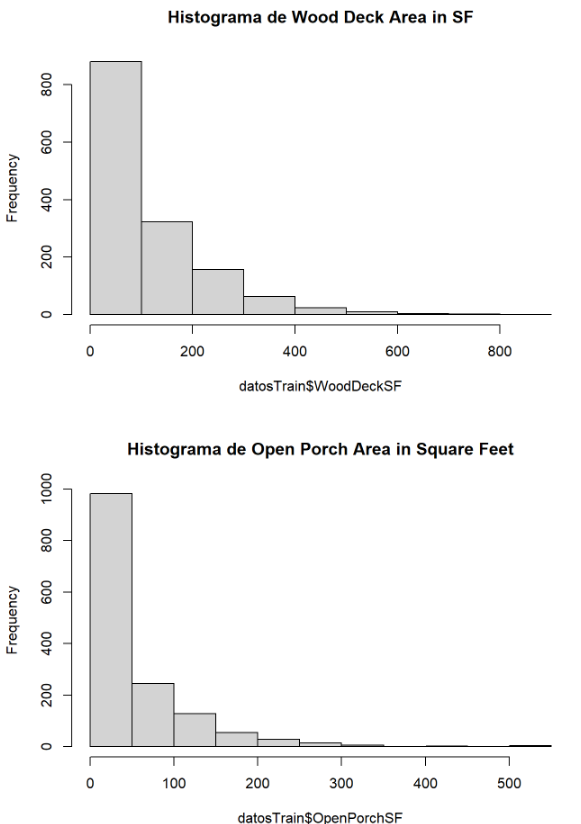


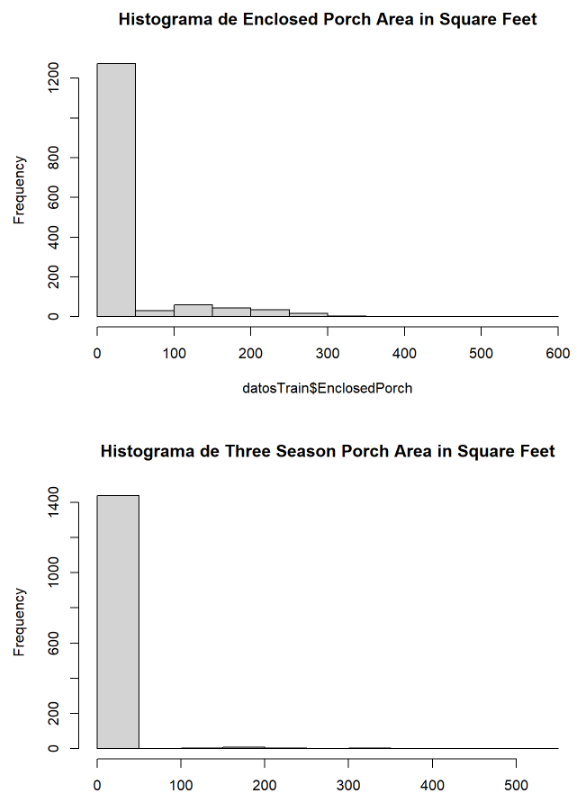
Se observa que Total Rooms Above Grade sigue una distribución normal

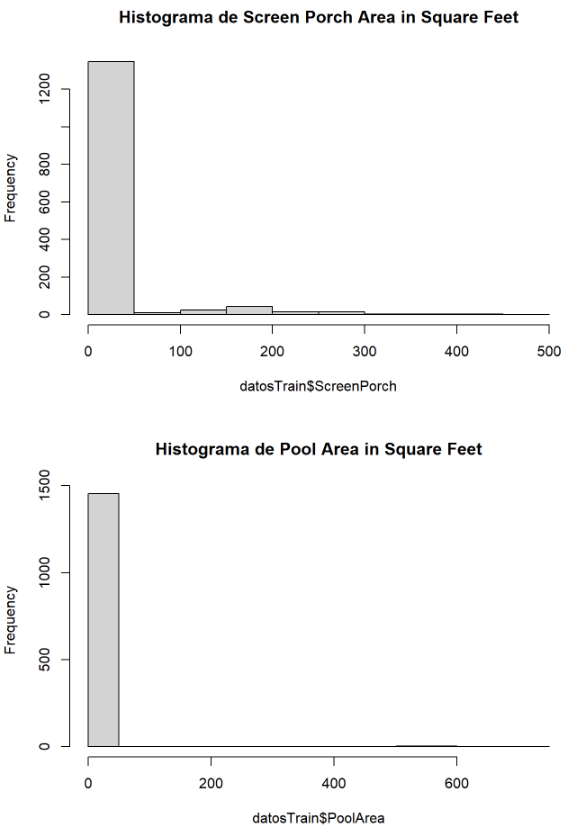


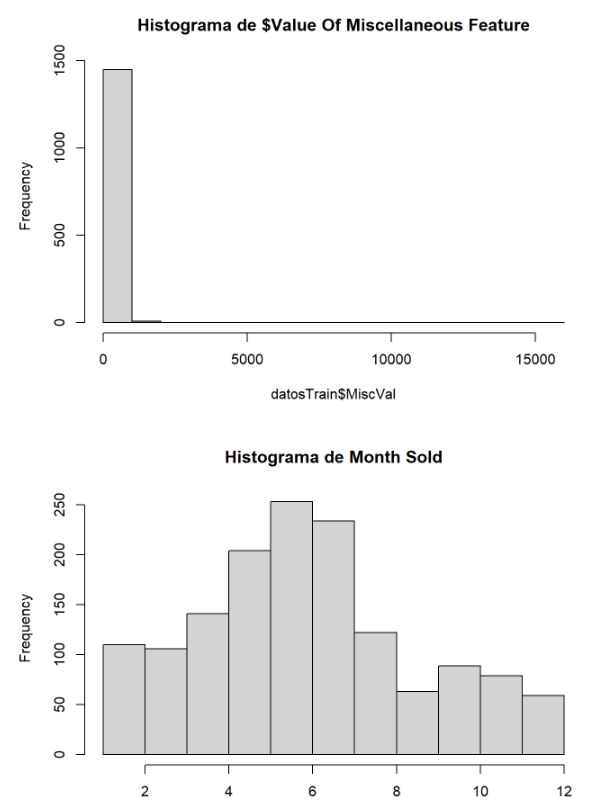


Ambos Size Of Garage in Car Capacity y Size Of Garage in Square Feet siguen una distribución normal.

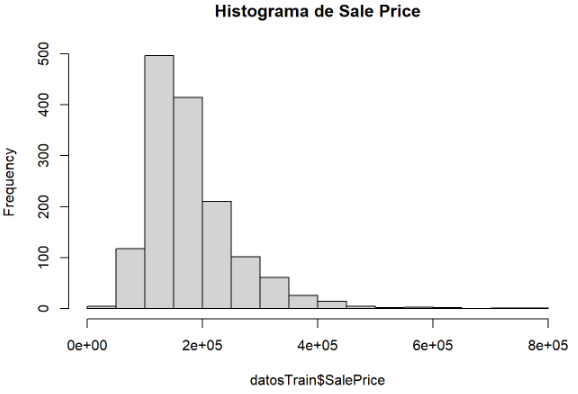








Sorprendentemente el mes vendido también parece tener una distribución normal.

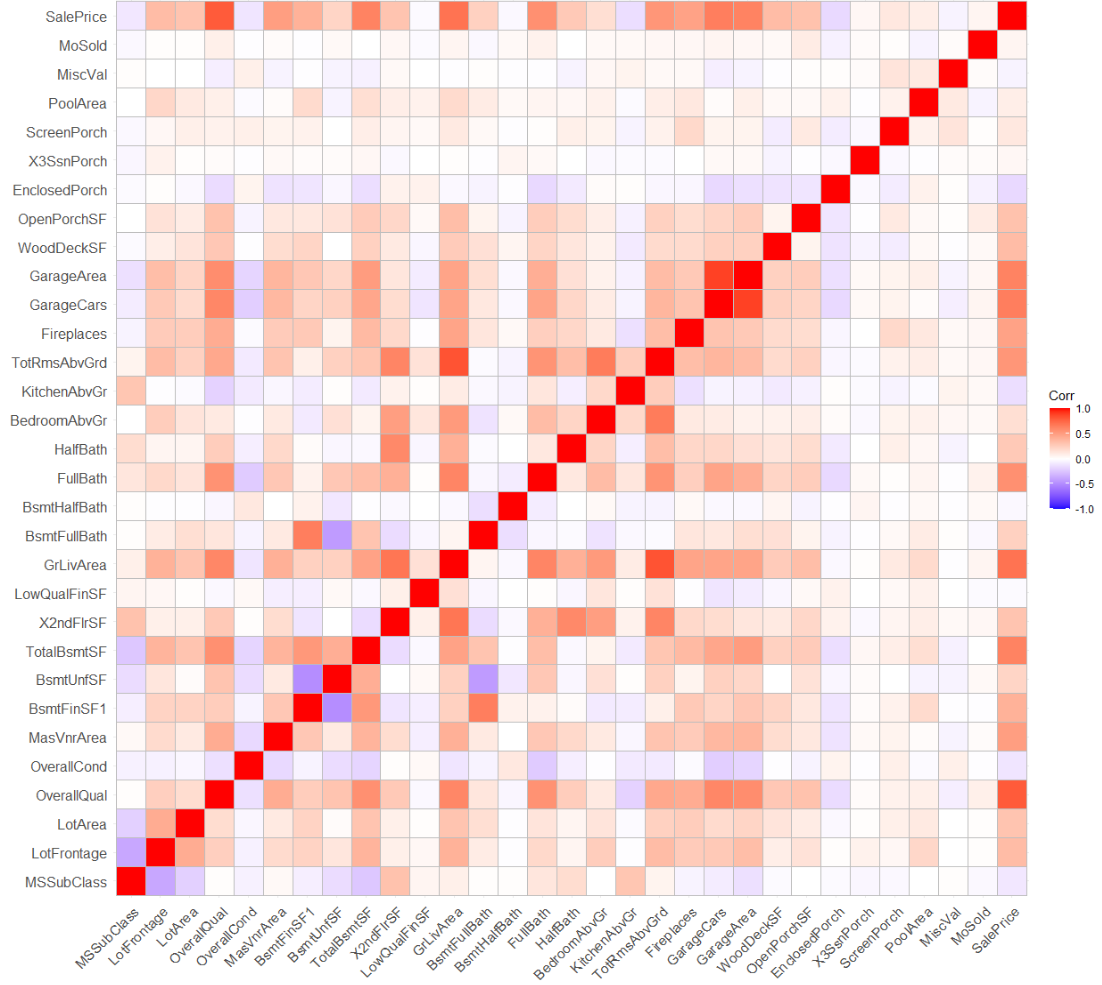


SalePrice también tiene una distribución normal.

Ya con los histogramas y las tablas de frecuencia, se pueden entender más los datos y se pudo observar que los datos con normalidad eran los siguientes: MonthSold, SalePrice, GarageArea, GarageCars, BedroomAbvGr, TotRmsAbvGrd, GrLivArea, X1stFlrSF, TotalBsmtSF, OverallQual, OverallCond, LotFrontage. Para un total de 12 variables.

**4. Aísle las variables numéricas de las categóticas, haga un análisis de correlación entre las mismas.**

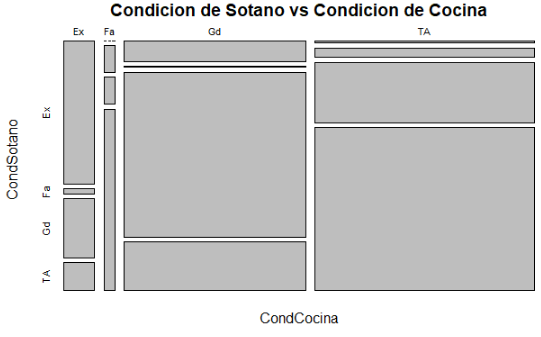
Se hizo una matriz de correlación con las variables numéricas y también se sacó el determinante de la misma el cuál es de: 1.1316e-07, que es básicamente 0. Esto indica que hay multicolinealidad entre las varibles y que más adelante será útil hacer un PCA, ya que para que un PCA sea efectivo, hay que tener correlacion de variables.

La matriz de correlacion cruda está en el RMarkDown, pero también se hizo un mapa de calor de correlación para verlo de una manera más facil.

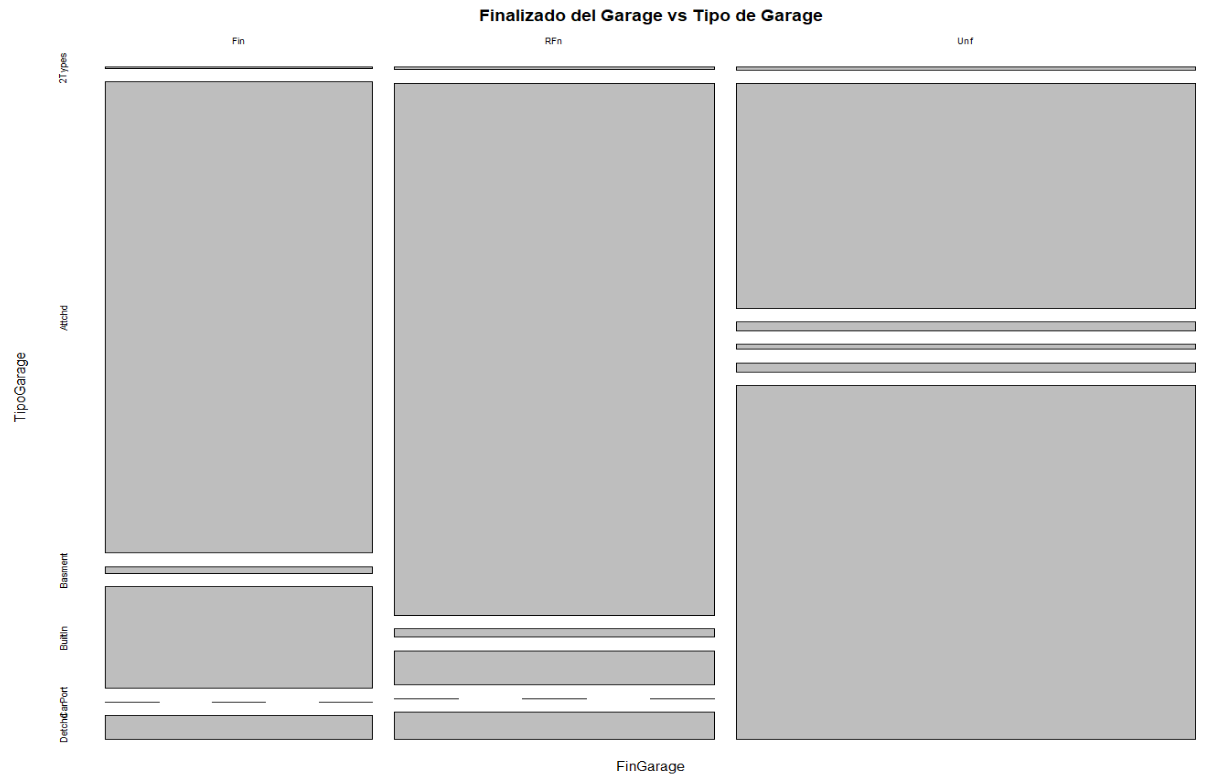
**­**

**5. Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos**

Como anteriormente, en el inciso 3, se muestran todas las tablas de frecuencia de las variables categóricas, ahora se buscará si hay relación entre algunos pares de variables que nos parezcan interesantes. Se comprobará la correlación con el algoritmo de Cramer-v



Se encontró una correlación moderada en la condición de las cocinas y la condición de los sótanos, ya que podemos observar que la mayoría de los sótanos excelentes (Ex) también son cocinas excelentes. El valor de Cramer-V es de 0.4214.

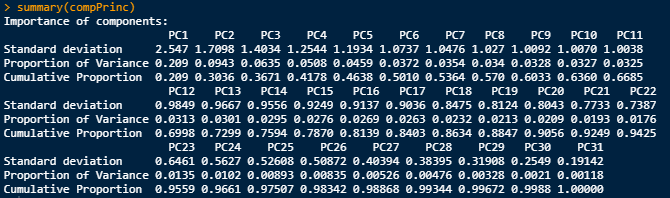


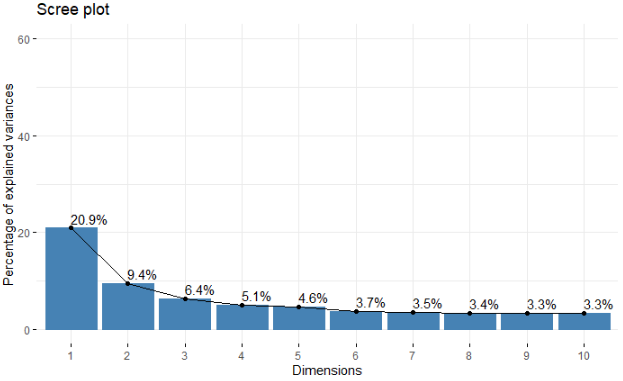
También se encontró una correlación moderada entre el finalizado del interior del garaje y el tipo del garaje ya que la mayoría de garajes sin finalizar (Unf) son garajes separados de la casa (Detch) y la mayoría de los que están junto a la casa o son parte de la casa (Attchd) estan Finalizados (Fin). Se obtuvo un valor Cramer-V de 0.4564.

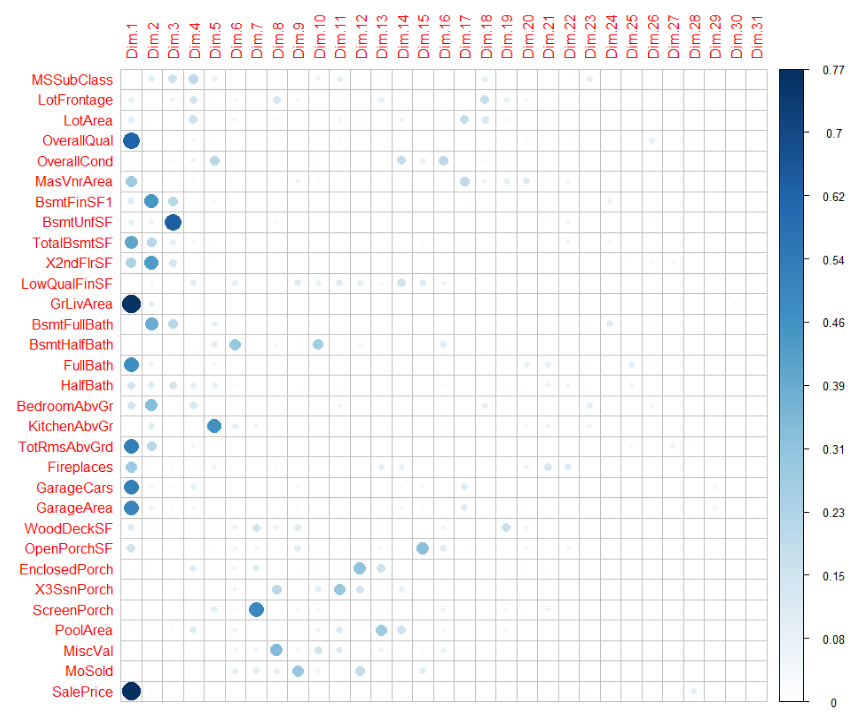
**6. Estudie si es conveniente hacer un Análisis de Componentes Principales. Recuerde que puede usar el índice KMO y el test de esfericidad de Bartlett. Haga un análisis de componentes principales con las variables numéricas, discuta los resultados e interprete los componentes.**

Para comenzar, nos podemos dar una idea de que sí es conveniente, ya que la determinante de la matriz de correlación era casi 0. Sin embargo, también se usó el índice de KMO y el test de esfericidad de Bartlett y el resultado fue de 0.78349 y 18919, respectivamente. Estos valores nos dicen que existe una aceptable adecuación muestral y es objetivamente útil hacer PCA.



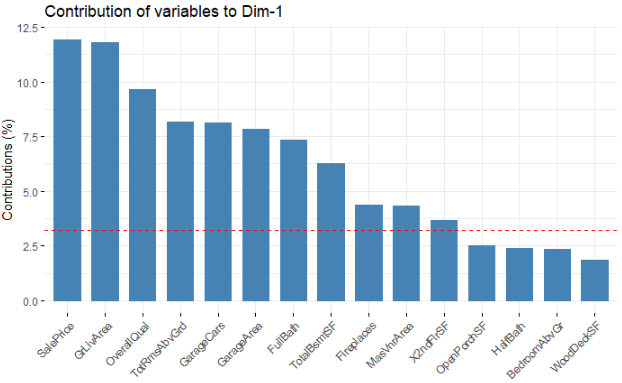
Aquí está el resumen de los componentes principales.

Se hizo un screeplot y como se puede ver, el porcentaje deja de bajar significantemente en la dimensión número 3 (6.4%), por lo que podemos inferir que podríamos representar el data set con 3 componentes.

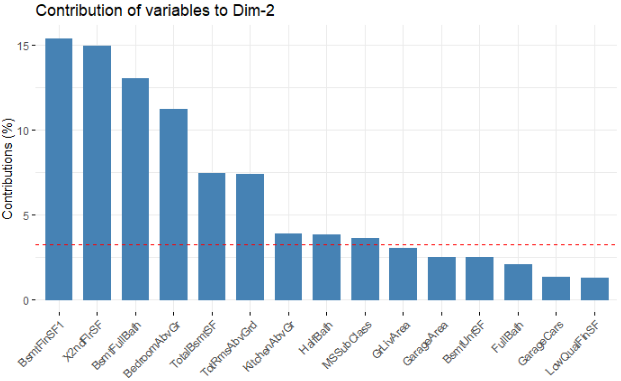
También se hizo el gráfico de la representación de cada variable en cada componente y podemos observar que efectivamente solamente 3 componentes es correcto, ya que después del tercero ya no hay una representación significativa de alguna variable

A continuación, se muestra la interpretación de los tres componentes principales:

Componente 1: **Calidad de la casa en general**

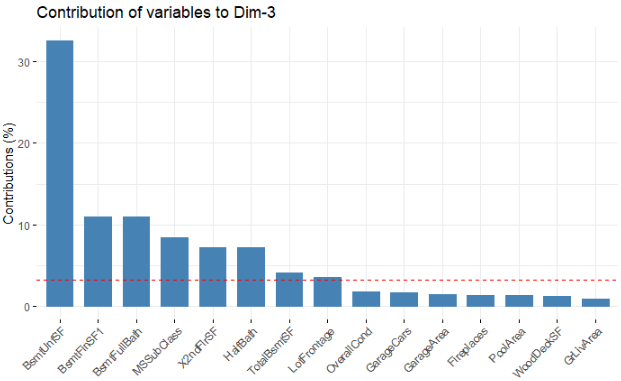


Las variables que mejor están representadas en el componente 1 son: SalePrice, GrLivArea, OverallQual, TotRmsAbvGrd, GarageCars, GarageArea, FullBath, TotalBsmtSF, FirePlaces, MasVnrArea, X2ndFlrSF. Al observar las primeras tres, las cuales son el precio de venta, el area en pies cuadrados de vivienda y la calidad en general. Se podría tomar este componente como la calidad de la casa en general, ya que por lo general una casa grande y que es cara, va a ser una casa de buena calidad.

Componente 2: **Calidad de áreas fuera del piso principal.**

Este componente esta principalmente representado por el finalizado del sótano (BsmtFinSF1), el área del segundo piso (X2ndFlrSF), la cantidad de baños completos del sótano (BsmtFullBath) y la cantidad de cuartos sin incluir los del sótano. No todas las casas tienen sótano o segundo piso, entonces este componente puede ser representado como calidad de áreas fuera del piso principal.

Componente 3: **Calidad del sótano**



El componente 3 está representado principalmente por el área sin finalizar del sótano en pies cuadrados (BsmtUnfSF) y el finalizado del sótano. No queda duda que este componente puede ser representado como calidad del sótano.

**7. Obtenga reglas de asociación interesantes del dataset. Discuta sobre el nivel de confianza y soporte.**

Se trabajaron solo con los datos cualitativos, así como se vio en clase. El nivel de confianza y soporte se colocaron más altos que en la clase para poder tener las reglas más interesantes. Se tuvo que trabajar por columnas (no todas en el mismo set ya que eran demasiadas columnas y R no lo soportaba). Como el dataset es tan extenso se crean una cantidad grande de reglas.

