Homework 4 - Data Wrangling in R

Daniel Carpenter & NAME HERE

September 2022

Table of contents

Packages

```
library(tidyverse)
```

1 - Data Quality Report

1 (a) - Read data

```
# Read data
housingData <- read_csv('housingData.csv')
```

```
Rows: 1000 Columns: 74
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (38): MSZoning, Alley, LotShape, LandContour, LotConfig, LandSlope, Neig...
dbl (36): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# create three new variables
housingData <- housingData %>%
  dplyr::mutate(age             = YrSold - YearBuilt,
                ageSinceRemodel = YrSold - YearRemodAdd,
                ageofGarage     = YrSold - GarageYrBlt
                )
```

1 (b) - Numeric Housing Tibble

- Create a tibble named `housingNumeric` which contains all of the numeric variables from the original data.
- use the `dplyr::select` command along with the `is.numeric` function to complete this task.

```
# Convert df to a tibble
housingNumeric <- as_tibble(housingData) %>%
```

```
  # Only select numeric data
  # note would usually use command select_if(is.numeric)
  select(where(is.numeric))
```

1 (c) - Factor Housing Tibble

- create a tibble named `housingFactor` which contains all of the character variables from the original data.

```
housingFactor <- as_tibble(housingData) %>%

  # Change all character variables to factors
  # Keep only the changed data. Implicitly keeping only factor (prev. char vars)
  transmute_if(is.character, as.factor)
```

1 (d) - Use Glimpse

```
# NOT RUN
# glimpse(housingNumeric)
# glimpse(housingFactor)
```

1 (e) - Get Q1 and Q3

- create our own user-defined functions for extracting only first and 3rd quantile
- Explanation: Gets the quantiles of a vector using quantile function, but we use the `[]` brackets to retrieve the 2nd or 4th objects in the vector, which are `Q1` and `Q3`

```
Q1 <- function(x,na.rm=TRUE) {
  quantile(x,na.rm=na.rm)[2]
}
Q3 <- function(x,na.rm=TRUE) {
  quantile(x,na.rm=na.rm)[4]
}
```

1 (f) - Vectorized Summary Stats

- Function that will help apply several summary statistics to our data all at once
- Contains vector of functions with default to not evaluate if `na`

```
# Vector of functions
myNumericSummary <- function(x){
  c(length(x), n_distinct(x), sum(is.na(x)), mean(x, na.rm=TRUE),
  min(x,na.rm=TRUE), Q1(x,na.rm=TRUE), median(x,na.rm=TRUE), Q3(x,na.rm=TRUE),
  max(x,na.rm=TRUE), sd(x,na.rm=TRUE))
}

# Name of each functions within the vector
statNames <- c('n', 'unique', 'missing', 'mean', 'min', 'Q1', 'median', 'Q3', 'max', 'sd')
```

1 (g) - Apply Summary Stats

- Apply summary stats function with `summarize` function

```
numericSummary <- housingNumeric %>%

  # Apply vector of functions using summarise
  summarise( across( where(is.numeric), ~myNumericSummary(.x) ) )
```

1 (h) - Add Stats Names

- Combine original data set and the names of each summary statistic

```
numericSummary <- cbind(
  stat=c("n","unique","missing","mean","min","Q1","median","Q3","max","sd"),
  numericSummary
)

# glimpse(numericSummary) # uncomment to see effects
```

1 (i) - Pretty up data

Transform data to make it ready for output format

```
numericSummaryFinal <- numericSummary %>%
  pivot_longer("Id":"ageofGarage", names_to = "variable", values_to = "value") %>%
  pivot_wider(names_from = stat, values_from = value) %>%
  mutate(missing_pct = 100*missing/n,
         unique_pct = 100*unique/n) %>%
  select(variable, n, missing, missing_pct, unique, unique_pct, everything())
```

Show the output

```
library(knitr)
options(digits=3)
options(scipen=99)
numericSummaryFinal %>% kable()
```

variable	n	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
Id	1000	0	0.0	1000	100.0	500.500	1	251	500	750.2	1000	288.819
MSSubClass	1000	0	0.0	13	1.3	57.185	20	20	50	70.0	190	41.875
LotFrontage	1000	207	20.7	102	10.2	68.745	21	58	68	80.0	313	23.198
LotArea	1000	0	0.0	760	76.0	10424.881	1477	7500	9422	11423.5	215245	9940.619
OverallQual	1000	0	0.0	10	1.0	5.979	1	5	6	7.0	10	1.310
OverallCond	1000	0	0.0	8	0.8	5.638	2	5	5	6.0	9	1.114
YearBuilt	1000	0	0.0	108	10.8	1969.836	1875	1954	1971	1998.0	2009	29.119
YearRemodAdd	1000	0	0.0	61	6.1	1984.108	1950	1967	1992	2002.0	2010	20.116
MasVnrArea	1000	4	0.4	249	24.9	95.418	0	0	0	146.2	1600	177.318
BsmtFinSF1	1000	0	0.0	490	49.0	438.686	0	0	400	700.0	1880	405.837
BsmtFinSF2	1000	0	0.0	107	10.7	44.296	0	0	0	0.0	1127	150.493
BsmtUnfSF	1000	0	0.0	598	59.8	535.078	0	208	441	779.2	2153	417.944
TotalBsmtSF	1000	0	0.0	549	54.9	1018.060	0	793	962	1223.5	3206	403.641
X1stFlrSF	1000	0	0.0	581	58.1	1131.251	334	868	1060	1327.2	3228	350.862
X2ndFlrSF	1000	0	0.0	306	30.6	346.279	0	0	0	735.0	1872	426.395
LowQualFinSF	1000	0	0.0	15	1.5	4.991	0	0	0	0.0	528	45.295
GrLivArea	1000	0	0.0	664	66.4	1482.521	334	1111	1442	1735.0	4316	490.566
BsmtFullBath	1000	0	0.0	3	0.3	0.427	0	0	0	1.0	2	0.509
BsmtHalfBath	1000	0	0.0	2	0.2	0.059	0	0	0	0.0	1	0.236
FullBath	1000	0	0.0	4	0.4	1.529	0	1	2	2.0	3	0.531

variable	n	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
HalfBath	1000	0	0.0	3	0.3	0.384	0	0	0	1.0	2	0.501
BedroomAbvGr	1000	0	0.0	7	0.7	2.865	0	2	3	3.0	6	0.791
KitchenAbvGr	1000	0	0.0	3	0.3	1.041	1	1	1	1.0	3	0.203
TotRmsAbvGrd	1000	0	0.0	11	1.1	6.410	2	5	6	7.0	12	1.562
Fireplaces	1000	0	0.0	4	0.4	0.618	0	0	1	1.0	3	0.642
GarageYrBlt	1000	53	5.3	94	9.4	1976.938	1906	1960	1977	1999.0	2009	23.592
GarageCars	1000	0	0.0	5	0.5	1.720	0	1	2	2.0	4	0.714
GarageArea	1000	0	0.0	353	35.3	458.329	0	319	470	572.0	1356	197.780
WoodDeckSF	1000	0	0.0	226	22.6	94.555	0	0	0	168.0	857	127.144
OpenPorchSF	1000	0	0.0	169	16.9	43.610	0	0	22	64.0	547	61.915
EncPorchSF	1000	0	0.0	122	12.2	40.641	0	0	0	0.0	508	82.139
PoolArea	1000	0	0.0	3	0.3	1.224	0	0	0	0.0	648	27.403
MiscVal	1000	0	0.0	14	1.4	27.210	0	0	0	0.0	3500	190.707
MoSold	1000	0	0.0	12	1.2	6.207	1	4	6	8.0	12	2.626
YrSold	1000	0	0.0	5	0.5	2007.919	2006	2007	2008	2009.0	2010	1.318
SalePrice	1000	0	0.0	477	47.7	174560.68	87300	130000	160000	205000.0	755000	69329.319
age	1000	0	0.0	115	11.5	38.083	1	10	37	55.0	135	29.109
ageSinceRemodel	1000	0	0.0	61	6.1	23.811	0	6	16	41.2	60	20.033
ageofGarage	1000	53	5.3	97	9.7	30.973	0	9	30	48.0	102	23.563

1 (j) - Factor Data Report

```
# TODO
```

2 - **Transformation**

```
housingData
```

```
# A tibble: 1,000 x 77
      Id MSSubCl~1 MSZon~2 LotFr~3 LotArea Alley LotSh~4 LandC~5 LotCo~6 LandS~7
   <dbl>     <dbl> <chr>     <dbl>   <dbl> <chr> <chr>   <chr>   <chr>   <chr>
 1     1        20 RL           NA   11000 <NA>  IR1     Lvl     CulDSac Gtl
 2     2        20 RL           NA   36500 <NA>  IR1     Low     Inside  Mod
 3     3        20 RL           57    9764 <NA>  IR1     Lvl     other   Gtl
 4     4        70 RL           NA    7500 <NA>  IR1     Bnk     Inside  Gtl
 5     5        20 RL           80    9200 <NA>  Reg     Lvl     Inside  Gtl
 6     6        60 RL           72   11317 <NA>  Reg     Lvl     Inside  Gtl
 7     7        20 RL           80    8480 <NA>  Reg     Lvl     Corner  Gtl
 8     8        70 RM           65   11700 Pave  IR1     Lvl     Corner  Gtl
 9     9        60 RL           80    9760 <NA>  Reg     Lvl     Inside  Mod
10    10        60 RL           93   10261 <NA>  IR1     Lvl     Inside  Gtl
# ... with 990 more rows, 67 more variables: Neighborhood <chr>,
#   Condition1 <chr>, BldgType <chr>, HouseStyle <chr>, OverallQual <dbl>,
#   OverallCond <dbl>, YearBuilt <dbl>, YearRemodAdd <dbl>, RoofStyle <chr>,
#   Exterior1st <chr>, Exterior2nd <chr>, MasVnrType <chr>, MasVnrArea <dbl>,
#   ExterQual <chr>, ExterCond <chr>, Foundation <chr>, BsmtQual <chr>,
#   BsmtCond <chr>, BsmtExposure <chr>, BsmtFinType1 <chr>, BsmtFinSF1 <dbl>,
#   BsmtFinType2 <chr>, BsmtFinSF2 <dbl>, BsmtUnfSF <dbl>, ...
```