Dhir Thacker | 17070122019 | CSE-1

SD lab assignment 7 (R-assignment 3)

Q1) What is the need of correlation analysis?

A1) Correlation analysis is used to determine and evaluate the relationship between two or more variables. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1 where +1 indicates perfect positive relation and -1 for negative or inverse relation. Pearson correlation, Kendall rank correlation, Spearman correlation, and the PointBiserial correlation are a few examples of correlation analysis techniques

Q2) Discuss one real world scenario where correlation helps to take decision.

A2) Take an example of a celestial body like a Star. We can correlate its size and the gravitational force it exhibits on nearby objects by calculating the correlation coefficient. In this case, it is obvious to expect a coefficient that tends to +1. A negative coefficient can be seen by considering the distance from the star, the greater the distance, lesser the gravitational force.

Q3) Write about function in R to compute correlation.

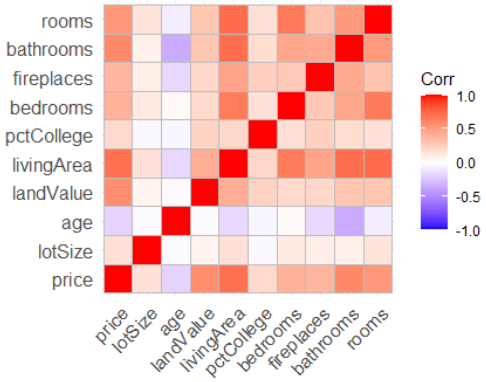
A3) A simple cor(x) can be used which returns the correlation coefficient. This function also allows us to choose the method of correlation as well as an option to handle missing data.

To test the relation between pairs, cor.test() is used. It returns both the correlation coefficient and the significance level (p-value) of the correlation.

Q4) Which kind of visualization (chart/graph) is suitable to represent correlation analysis graphically.

A4) A Scatter Chart is commonly used to visualize the correlation and distribution between two variables. These charts can show clusters that can help in detecting outliers.

Another widely used map is the Heat Map. This is more suitable when we want to visualize the correlation between multiple variables within a single plot



Q.5 Write a note on regression analysis.

A5) Regression analysis is a statistical method for estimating the relationships between variables in a data set. A dependent variable and one or more independent variables are required to perform the analysis. Generally, a dependent variable is what we wish to predict based on the values/parameters provided by the independent variables. Using this, a linear regression model can be generated which can help in predicting the outcome based on independent variable parameters.

There are 3 main types of regression analysis –

1. **Linear**
   Dependent and a single independent variable show a linear relationship

2. **Multilinear**
   Similar to linear, but the difference here is that multiple independent variables are involved

3. **Non-linear**
   Commonly used for complicated datasets where the variables show a nonlinear. Unlike linear models which show a simple curve line, non-linear models can have a variety of curve lines indicating the complexity of the data.

Correlation analysis in R

```r
# pierson correlation

# importing dataset
data(SaratogaHouses, package = "mosaicData")
View(SaratogaHouses)
```

r_assignment_3.R | df | r | SaratogaHouses

Filter

	price	lotSize	age	landValue	livingArea	pctCollege	bedrooms	fireplaces	bathrooms	rooms	heating	fuel	sewer	waterf
1	132500	0.09	42	50000	906	35	2	1	1.0	5	electric	electric	septic	No
2	181115	0.92	0	22300	1953	51	3	0	2.5	6	hot water/steam	gas	septic	No
3	109000	0.19	133	7300	1944	51	4	1	1.0	8	hot water/steam	gas	public/commercial	No
4	155000	0.41	13	18700	1944	51	3	1	1.5	5	hot air	gas	septic	No
5	86060	0.11	0	15000	840	51	2	0	1.0	3	hot air	gas	public/commercial	No
6	120000	0.68	31	14000	1152	22	4	1	1.0	8	hot air	gas	septic	No
7	153000	0.40	33	23300	2752	51	4	1	1.5	8	hot water/steam	oil	septic	No
8	170000	1.21	23	14600	1662	35	4	1	1.5	9	hot air	oil	septic	No
9	90000	0.83	36	22200	1632	51	3	0	1.5	8	electric	electric	septic	No
10	122900	1.94	4	21200	1416	44	3	0	1.5	6	hot air	gas	none	No
11	325000	2.29	123	12600	2894	51	7	0	1.0	12	hot air	oil	septic	No
12	120000	0.92	1	22300	1624	51	3	0	2.0	6	hot air	gas	septic	No
13	85860	8.97	13	4800	704	41	2	0	1.0	4	electric	electric	septic	No
14	97000	0.11	153	3100	1383	57	3	0	2.0	5	hot water/steam	gas	public/commercial	No
15	127000	0.14	9	300	1300	41	3	0	1.5	8	hot air	oil	septic	No
16	89900	0.00	88	2500	936	57	3	0	1.0	4	hot water/steam	gas	public/commercial	No
17	155000	0.13	9	300	1300	41	3	0	1.5	7	hot air	oil	septic	No
18	253750	2.00	0	49800	2816	71	4	1	2.5	12	hot air	gas	none	No

Showing 1 to 18 of 1,728 entries, 16 total columns

```r
# selecting numeric variables
df <- dplyr::select_if(SaratogaHouses, is.numeric)
```

r_assignment_3.R* | df | r | SaratogaHouses

Filter

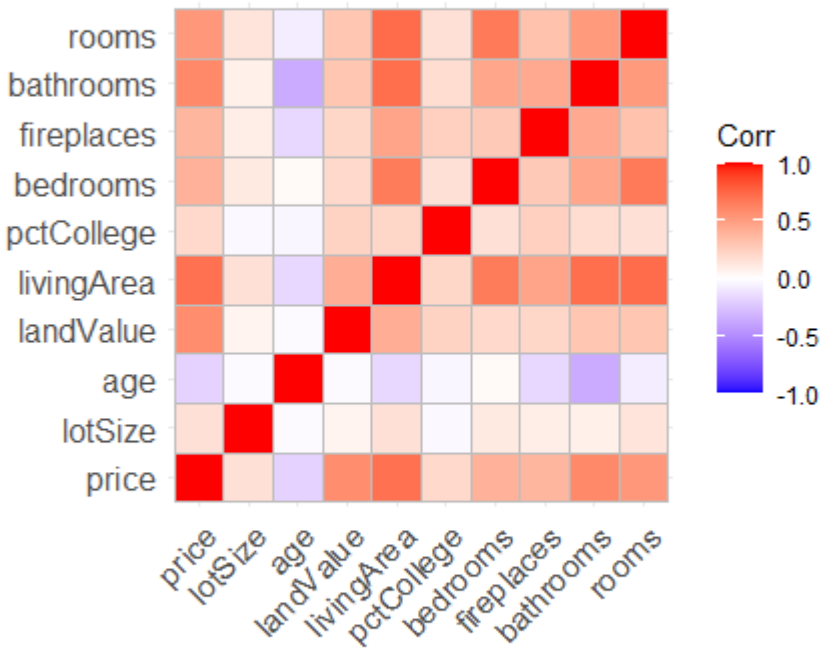
	price	lotSize	age	landValue	livingArea	pctCollege	bedrooms	fireplaces	bathrooms	rooms
1	132500	0.09	42	50000	906	35	2	1	1.0	5
2	181115	0.92	0	22300	1953	51	3	0	2.5	6
3	109000	0.19	133	7300	1944	51	4	1	1.0	8
4	155000	0.41	13	18700	1944	51	3	1	1.5	5
5	86060	0.11	0	15000	840	51	2	0	1.0	3
6	120000	0.68	31	14000	1152	22	4	1	1.0	8
7	153000	0.40	33	23300	2752	51	4	1	1.5	8
8	170000	1.21	23	14600	1662	35	4	1	1.5	9
9	90000	0.83	36	22200	1632	51	3	0	1.5	8
10	122900	1.94	4	21200	1416	44	3	0	1.5	6
11	325000	2.29	123	12600	2894	51	7	0	1.0	12
12	120000	0.92	1	22300	1624	51	3	0	2.0	6
13	85860	8.97	13	4800	704	41	2	0	1.0	4
14	97000	0.11	153	3100	1383	57	3	0	2.0	5
15	127000	0.14	9	300	1300	41	3	0	1.5	8
16	89900	0.00	88	2500	936	57	3	0	1.0	4
17	155000	0.13	9	300	1300	41	3	0	1.5	7
18	253750	2.00	0	49800	2816	71	4	1	2.5	12

Showing 1 to 19 of 1,728 entries, 10 total columns

```r
# calculating the correlations
r <- cor(df, use="complete.obs")
round(r, 2)
```

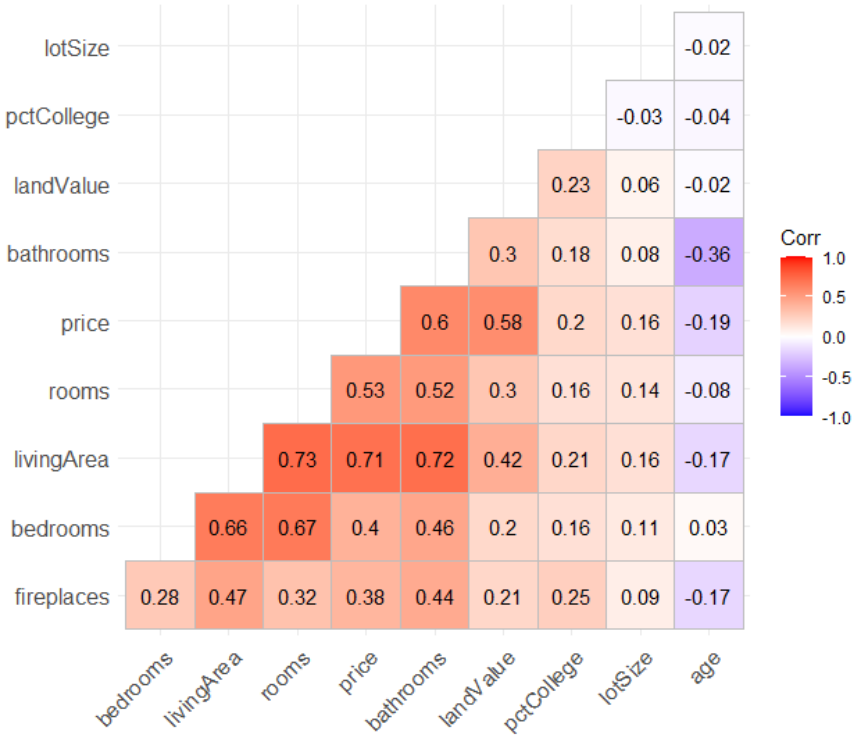
```
> # calculating the correlations
> r <- cor(df, use="complete.obs")
> round(r, 2)
            price lotSize   age landValue livingArea pctCollege bedrooms fireplaces bathrooms rooms
price        1.00    0.16 -0.19      0.58       0.71       0.20     0.40       0.38      0.60  0.53
lotSize      0.16    1.00 -0.02      0.06       0.16      -0.03     0.11       0.09      0.08  0.14
age         -0.19   -0.02  1.00     -0.02      -0.17      -0.04     0.03      -0.17     -0.36 -0.08
landValue    0.58    0.06 -0.02      1.00       0.42       0.23     0.20       0.21      0.30  0.30
livingArea   0.71    0.16 -0.17      0.42       1.00       0.21     0.66       0.47      0.72  0.73
pctCollege   0.20   -0.03 -0.04      0.23       0.21       1.00     0.16       0.25      0.18  0.16
bedrooms     0.40    0.11  0.03      0.20       0.66       0.16     1.00       0.28      0.46  0.67
fireplaces   0.38    0.09 -0.17      0.21       0.47       0.25     0.28       1.00      0.44  0.32
bathrooms    0.60    0.08 -0.36      0.30       0.72       0.18     0.46       0.44      1.00  0.52
rooms        0.53    0.14 -0.08      0.30       0.73       0.16     0.67       0.32      0.52  1.00
>
```

```r
library(ggplot2)
library(ggcorrplot)
# dark red - strong positive correlations
# dark blue - strong negative correlations
# white - no correlations
ggcorrplot(r)
```




```r
ggcorrplot(r,
           hc.order = TRUE,
           type = "lower",
           lab = TRUE)

# hc.order = TRUE reorders variables, placing similar correlation
# patterns together.
# type = "lower" plots the lower portion of the correlation matrix.
# lab = TRUE overlays the correlation coefficients (as text) on the plot.
```


lotSize
pctCollege
landValue
bathrooms
price
rooms
livingArea
bedrooms
fireplaces
-0.02
-0.03 -0.04
0.23 0.06 -0.02
0.3 0.18 0.08 -0.36
0.6 0.58 0.2 0.16 -0.19
0.53 0.52 0.3 0.16 0.14 -0.08
0.73 0.71 0.72 0.42 0.21 0.16 -0.17
0.66 0.67 0.4 0.46 0.2 0.16 0.11 0.03
0.28 0.47 0.32 0.38 0.44 0.21 0.25 0.09 -0.17
bedrooms
livingArea
rooms
price
bathrooms
landValue
pctCollege
lotSize
age
Corr
1.0
0.5
0.0
-0.5
-1.0