Week 4 Missing Data and Outliers Analysis

Nitin

September 24, 2020

Read and Analyze Data

Leat us analisis diffrence between when we read file as csv or xlsx

```
library(readxl)
XlData <- read_excel("data.xlsx")
str(XlData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    361 obs. of  25 variables:
##  $ Month                        : POSIXct, format: "1990-04-01" "1990-05-01" ...
##  $ Coarse wool Price            : num  482 447 441 418 418 ...
##  $ Coarse wool price % Change   : chr  "-" "-7.2700000000000001E-2" "-1.4E-2" "-5.11E-2" ...
##  $ Copra Price                  : num  236 234 216 205 198 196 198 236 237 233 ...
##  $ Copra price % Change         : chr  "-" "-8.5000000000000006E-3" "-7.6899999999999996E-2" "-5.0900
##  $ Cotton Price                 : num  1.83 1.89 1.99 2.01 1.79 1.79 1.79 1.82 1.85 1.85 ...
##  $ Cotton price % Change        : chr  "-" "3.2800000000000003E-2" "5.2900000000000003E-2" "1.01E-2"
##  $ Fine wool Price              : num  1072 1057 898 896 951 ...
##  $ Fine wool price % Change     : chr  "-" "-1.35E-2" "-0.15029999999999999" "-2.7000000000000001E-3"
##  $ Hard log Price               : num  161 173 182 188 186 ...
##  $ Hard log price % Change      : chr  "-" "7.2300000000000003E-2" "5.0999999999999997E-2" "3.4599999
##  $ Hard sawnwood Price          : num  550 492 495 486 488 ...
##  $ Hard sawnwood price % Change: chr  "-" "-0.1055" "7.1000000000000004E-3" "-1.9199999999999998E-2"
##  $ Hide Price                   : num  100 99.5 97.9 96.8 91.9 ...
##  $ Hide price % change          : chr  "-" "-5.4000000000000003E-3" "-1.5699999999999999E-2" "-1.17E-
##  $ Plywood Price                : num  312 350 374 378 365 ...
##  $ Plywood price % Change       : chr  "-" "0.12089999999999999" "6.8000000000000005E-2" "1.21E-2" ..
##  $ Rubber Price                 : num  0.84 0.85 0.85 0.86 0.88 0.9 0.9 0.9 0.88 0.87 ...
##  $ Rubber price % Change        : chr  "-" "1.1900000000000001E-2" "0" "1.18E-2" ...
##  $ Softlog Price                : num  121 124 129 124 130 ...
##  $ Softlog price % Change       : chr  "-" "0.03" "4.1599999999999998E-2" "-4.0300000000000002E-2" ..
##  $ Soft sawnwood Price          : num  219 213 200 210 208 ...
##  $ Soft sawnwood price % Change: chr  "-" "-2.63E-2" "-6.0999999999999999E-2" "5.0299999999999997E-2
##  $ Wood pulp Price              : num  829 843 831 799 819 ...
##  $ Wood pulp price % Change     : chr  "-" "1.5900000000000001E-2" "-1.32E-2" "-3.9100000000000003E-2
```

Create new data frame with columns of our Intrest.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages ---------------------------------------------------------------------------- t
```

```
## v ggplot2 3.2.0       v purrr   0.2.5
## v tibble  2.1.3       v dplyr   0.8.0.1
## v tidyr   0.8.1       v stringr 1.3.1
## v readr   1.1.1       v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
## -- Conflicts ---------------------------------------------------------------------------------- tidyver
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df2 <- select(XlData, -c(3,5,7,9,11,13,15,17,19,21,23,25))
str(df2)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    361 obs. of  13 variables:
##  $ Month              : POSIXct, format: "1990-04-01" "1990-05-01" ...
##  $ Coarse wool Price  : num  482 447 441 418 418 ...
##  $ Copra Price        : num  236 234 216 205 198 196 198 236 237 233 ...
##  $ Cotton Price       : num  1.83 1.89 1.99 2.01 1.79 1.79 1.79 1.82 1.85 1.85 ...
##  $ Fine wool Price    : num  1072 1057 898 896 951 ...
##  $ Hard log Price     : num  161 173 182 188 186 ...
##  $ Hard sawnwood Price: num  550 492 495 486 488 ...
##  $ Hide Price         : num  100 99.5 97.9 96.8 91.9 ...
##  $ Plywood Price      : num  312 350 374 378 365 ...
##  $ Rubber Price       : num  0.84 0.85 0.85 0.86 0.88 0.9 0.9 0.9 0.88 0.87 ...
##  $ Softlog Price      : num  121 124 129 124 130 ...
##  $ Soft sawnwood Price: num  219 213 200 210 208 ...
##  $ Wood pulp Price    : num  829 843 831 799 819 ...
```

1. Describe missing data, provide summary of missing data, similar to the analysis in the Chapter 2 (table 3): Count of missing data/percent per variable, type of missing data (NA, null), total percent of missingness per dataset.

```
summary(df2)
```

```
##      Month                    Coarse wool Price  Copra Price
##  Min.   :1990-04-01 00:00:00   Min.   : 247.1    Min.   : 182
##  1st Qu.:1997-10-01 00:00:00   1st Qu.: 369.6    1st Qu.: 372
##  Median :2020-03-12 00:00:00   Median : 525.1    Median : 458
##  Mean   :2011-08-01 03:03:29   Mean   : 626.3    Mean   : 542
##  3rd Qu.:2020-08-05 00:00:00   3rd Qu.: 847.1    3rd Qu.: 714
##  Max.   :2020-12-19 00:00:00   Max.   :1391.5    Max.   :1503
##                                NA's   :34        NA's   :22
##   Cotton Price  Fine wool Price  Hard log Price  Hard sawnwood Price
##  Min.   :0.82   Min.   : 417.5   Min.   :133.3   Min.   :413.4
##  1st Qu.:1.29   1st Qu.: 646.3   1st Qu.:198.0   1st Qu.:573.5
##  Median :1.60   Median : 748.2   Median :253.0   Median :728.7
##  Mean   :1.64   Mean   : 850.1   Mean   :251.0   Mean   :708.0
##  3rd Qu.:1.85   3rd Qu.:1019.9   3rd Qu.:283.0   3rd Qu.:831.6
##  Max.   :5.06   Max.   :1865.4   Max.   :520.8   Max.   :973.6
##                 NA's   :34                       NA's   :34
##    Hide Price     Plywood Price    Rubber Price   Softlog Price
##  Min.   : 28.59   Min.   :312.4   Min.   :0.490   Min.   :119.3
##  1st Qu.: 69.50   1st Qu.:442.5   1st Qu.:0.860   1st Qu.:146.0
##  Median : 77.25   Median :505.0   Median :1.440   Median :160.4
##  Mean   : 78.57   Mean   :508.2   Mean   :1.656   Mean   :164.5
##  3rd Qu.: 86.00   3rd Qu.:570.8   3rd Qu.:2.060   3rd Qu.:180.2
##  Max.   :114.63   Max.   :751.8   Max.   :6.260   Max.   :260.0
```

```
##  NA's   :34                                      NA's   :34
##  Soft sawnwood Price Wood pulp Price
##  Min.   :183.6       Min.   :384.0
##  1st Qu.:277.6       1st Qu.:549.8
##  Median :295.0       Median :693.6
##  Mean   :291.1       Mean   :696.7
##  3rd Qu.:310.9       3rd Qu.:875.0
##  Max.   :372.6       Max.   :966.5
##  NA's   :34          NA's   :1
```

From summary of data set it could be observed that various columns have some na observation like in case of Coarse wool Price column has 34 na's, Copra Price has 22 na's, Fine wool Price has 34 na's, Hard sawnwood Price 34 na's, Hide Price 34 na'sa, Softlog Price 34 na's, Soft sawnwood Price 34 na's and Wood pulp Price has 1 na.

Count of missing data/percent per variable

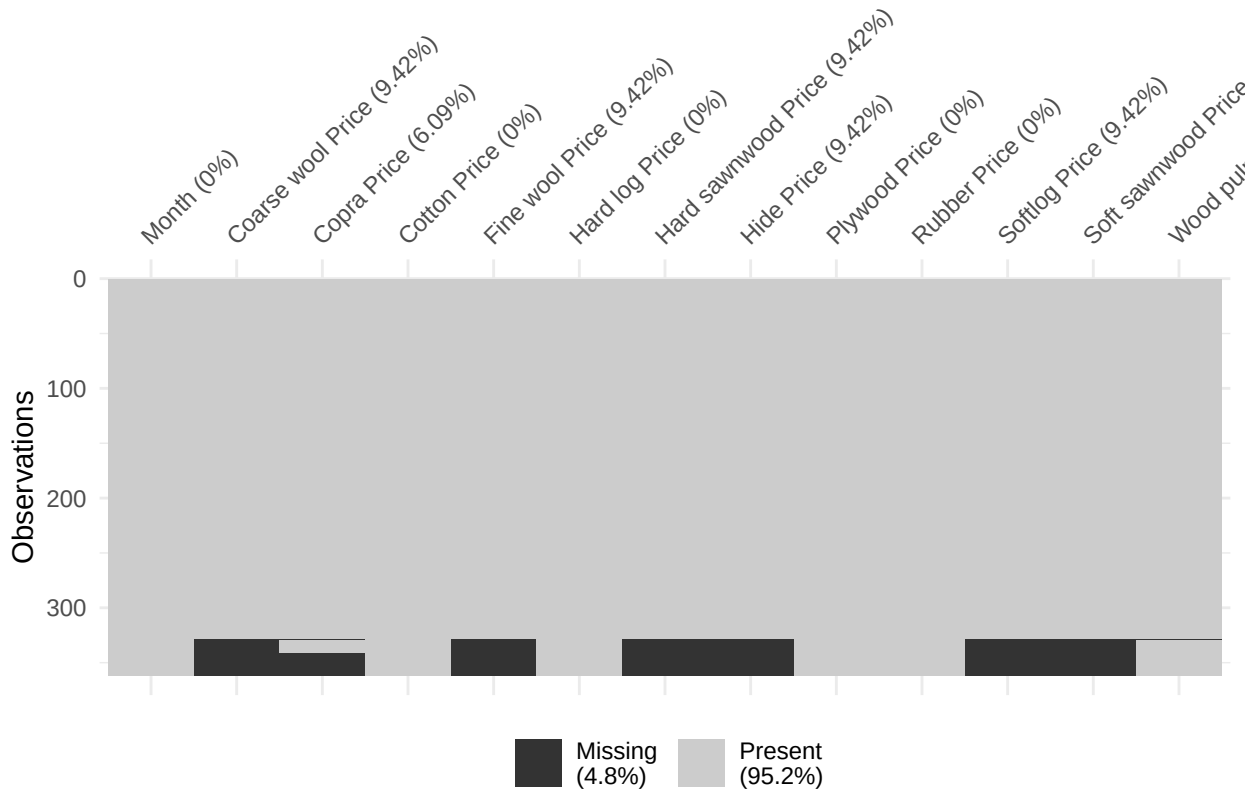
```
library(naniar)
miss_var_summary(df2)
```

```
## # A tibble: 13 x 3
##    variable            n_miss pct_miss
##    <chr>                <int>    <dbl>
##  1 Coarse wool Price       34    9.42
##  2 Fine wool Price         34    9.42
##  3 Hard sawnwood Price     34    9.42
##  4 Hide Price              34    9.42
##  5 Softlog Price           34    9.42
##  6 Soft sawnwood Price     34    9.42
##  7 Copra Price             22    6.09
##  8 Wood pulp Price          1    0.277
##  9 Month                    0    0
## 10 Cotton Price             0    0
## 11 Hard log Price           0    0
## 12 Plywood Price            0    0
## 13 Rubber Price             0    0
```

Table above displays missing data in percentage for all columns.

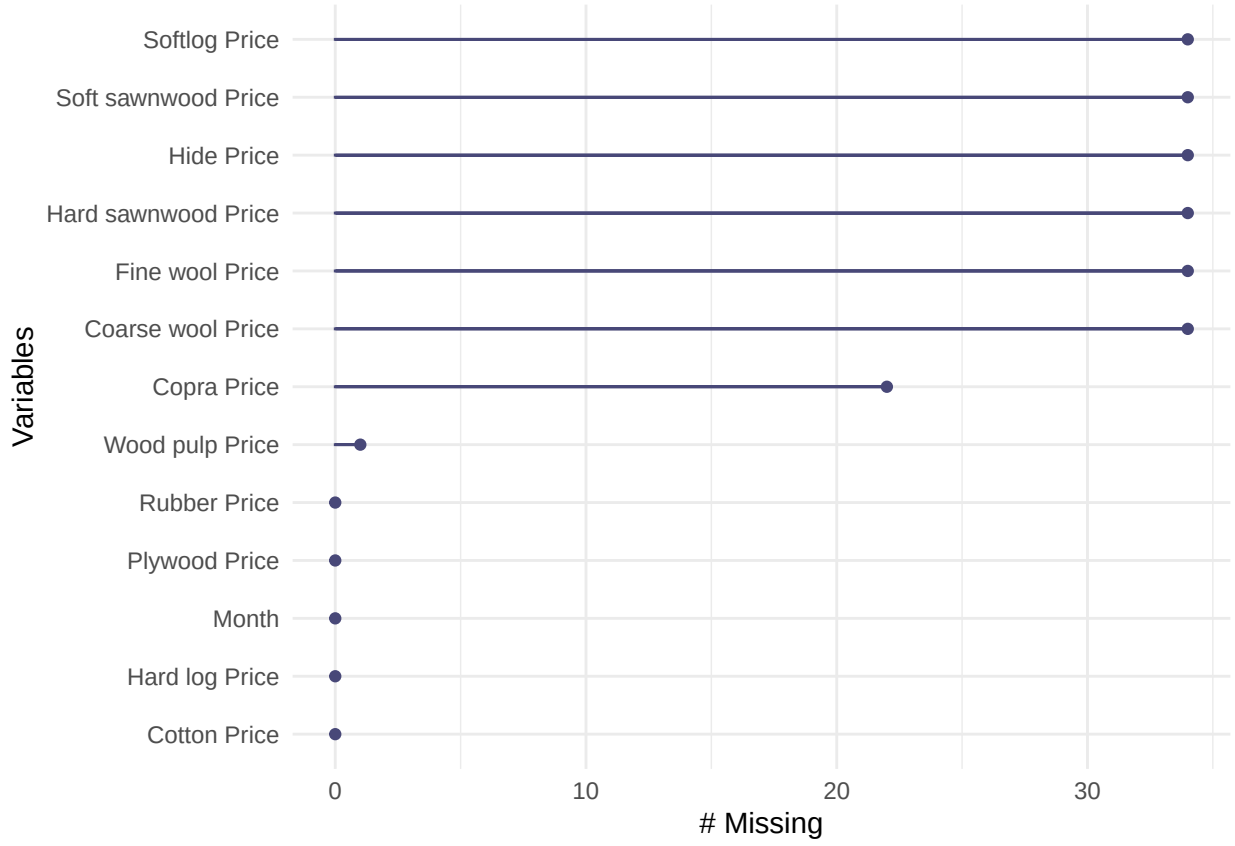
2. Plot visualization of missing data pattern.

```
vis_miss(df2, cluster = TRUE)
```



It could be observed from visualization for missing data it follows some pattern in missing observation. For most of the columns data is missing after observation number 340 onwards.

```
gg_miss_var(df2)
```



3. Describe if you have observed any patterns.

It could be observed that columns which are missing data it is after observation three hundred forty-one onwards.

4. Perform imputation, if needed: list-wise/pair-wise deletions, mean imputation, regression imputation etc. If imputation is not needed, explain why.

Since missing data is less than ten percent for all columns, we can ignore it or we can delete row for which data is missing.