R - Data cleaning and preprocessing

This assignment to exercise cleaning and preparing data for data mining. You are required to **code, run, and answer the following**:

1. Using R Studio (or any R compiler), install the following packages as follows:

```
pkgs <- sort(c('tidyverse', 'GGally', 'ggcorrplot',
    'plotly', 'factoextra', 'arules', 'seriation',
    'sampling', 'caret', 'proxy'))

pkgs_install <- pkgs[!(pkgs %in% installed.packages()[,"Package"])]
if(length(pkgs_install)) install.packages(pkgs_install)
```

2. Give a brief description of each package of ten packages above.

3. Load Iris dataset as follows:

```
library(tidyverse)
data(iris)
iris <- as_tibble(iris)
iris
```

4. Briefly describe Iris dataset.
5. Show the output of the above code and elaborate on the output.

```
print(iris, n = 3, width = Inf)
```

6. Show the output of the above code line and elaborate on the parameters and output.

```
summary(iris)
```

7. Show the output of the above code line and elaborate on the output.

```
iris %>% summarize_if(is.numeric, mean)
```

8. Show the output of the above code line and elaborate on the parameters and output.

```
library(GGally)
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
ggpairs(iris, aes(color = Species), progress = FALSE)
```

9. Show the output of the above code and elaborate on the parameters and output.

```
clean.data <- iris %>%
  drop_na() %>%
  unique()

summary(clean.data)
```

10.Show the output of the above code and elaborate on the output.

```
iris %>% group_by(Species) %>% summarize_all(mean)
iris %>% group_by(Species) %>% summarize_all(median)
```

11.Show the output of the above code and elaborate on the output.

```
sample(c("A", "B", "C"), size = 10, replace = TRUE)
```

12.Show the output of the above code line and elaborate on the output.

```
take <- sample(seq(nrow(iris)), size = 15)
take
iris[take, ]
```

13.Show the output of the above code and elaborate on the output.

```
set.seed(1000)
s <- iris %>% slice_sample(n = 15)
ggpairs(s, aes(color = Species), progress = FALSE)
```

14.Show the output of the above code and elaborate on the output.

```
library(sampling)
id2 <- strata(iris, stratanames = "Species",
              size = c(5,5,5), method = "srswor")
id2
s2 <- iris %>% slice(id2$ID_unit)

ggpairs(s2, aes(color = Species), progress = FALSE)
```

15.Show the output of the above code and elaborate on the output.

Dimensionality Reduction

Principal Components Analysis (PCA)

```
# library(plotly) # I don't load the package because it's namespace clashes with select in dplyr.
plotly::plot_ly(iris, x = ~Sepal.Length, y = ~Petal.Length, z = ~Sepal.Width,
      color = ~Species, size = 1) %>% plotly::add_markers()


pc <- iris %>% select(-Species) %>% as.matrix() %>% prcomp()
summary(pc)
plot(pc, type = "line")
str(pc)
```

16.Show the output of the above code and elaborate on the output.

```
iris_projected <- as_tibble(pc$x) %>% add_column(Species = iris$Species)

ggplot(iris_projected, aes(x = PC1, y = PC2, color = Species)) +
  geom_point()
```

17.Show the output of the above code and elaborate on the output.

```
library(factoextra)
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
fviz_pca(pc)
fviz_pca_var(pc)
```

18.Show the output of the above code and elaborate on the output.

Discretize Features

```
ggplot(iris, aes(x = Petal.Width)) + geom_histogram(binwidth = .2)
```

19.Show the output of the above code and elaborate on the output.

Proximities: Similarities and Distances

Minkowsky Distances

The Minkowsky distance is a family of metric distances including Euclidean and Manhattan distance. To avoid one feature to dominate the distance calculation, scaled data is typically used. We select the first 5 flowers for this example.

```
iris_sample <- iris.scaled %>% select(-Species) %>% slice(1:5)
iris_sample
dist(iris_sample, method = "euclidean")
dist(iris_sample, method = "manhattan")
dist(iris_sample, method = "maximum")
```

20.Show the output of the above code and elaborate on the output.

Relationships Between Features

Correlation

```
cc <- iris %>% select(-Species) %>% cor()
cc

with(iris, cor(Petal.Length, Petal.Width))
with(iris, cor.test(Petal.Length, Petal.Width))

ggplot(iris, aes(Petal.Length, Petal.Width)) +
  geom_point() +
  geom_smooth(method = "lm")
```

21.Show the output of the above code and elaborate on the output.