FIT3152 Data analytics. Tutorial 01: Introduction to R. Review of basic statistics Solutions

Pre-tutorial activities:

1. Read the data from the file "Ped_Count_December_2021.csv" into a data frame named "December". This contains hourly pedestrian activity at different sensor locations during December 2021. It was created using the City of Melbourne's automated pedestrian counting system. Run the command "December[3:83] = lapply(December[3:83], as.numeric)" to convert relevant column values to numeric. Answer the following questions making use of appropriate commands and functions in R.
   Source: http://www.pedestrian.melbourne.vic.gov.au/#date=18-02-2022&time=8

Answer the following questions:

How many rows and columns are there in the data set?

```
> dim(December)
   744  83
#alternatively use ncol(), nrow(), str() etc.
```

How many sensor locations are there in the dataset?

There are 83 columns. Without the first two" Date" and "Hour" columns, there are 81 sensor locations for which pedestrian counts are available.

How many observations are there each day?

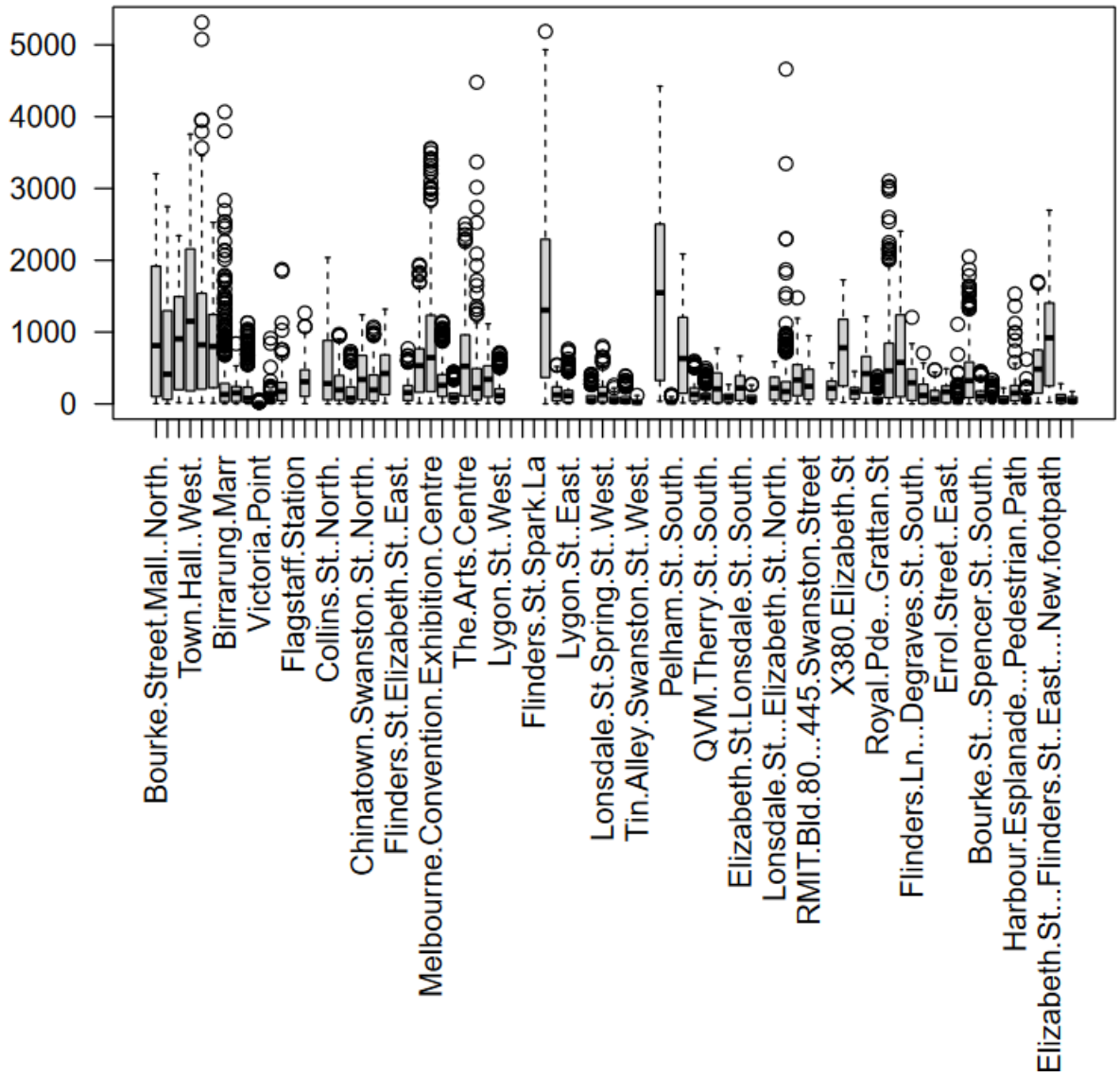
Given that the counts are given hourly. There are 24*81 =1944 observations per each day in the dataset. Note that there are missing values indicated as NA in multiple locations.

Calculate the average pedestrian count for each of the locations. Express your answers to two decimal places.

> round(colMeans(December[3:83], na.rm = TRUE),2)

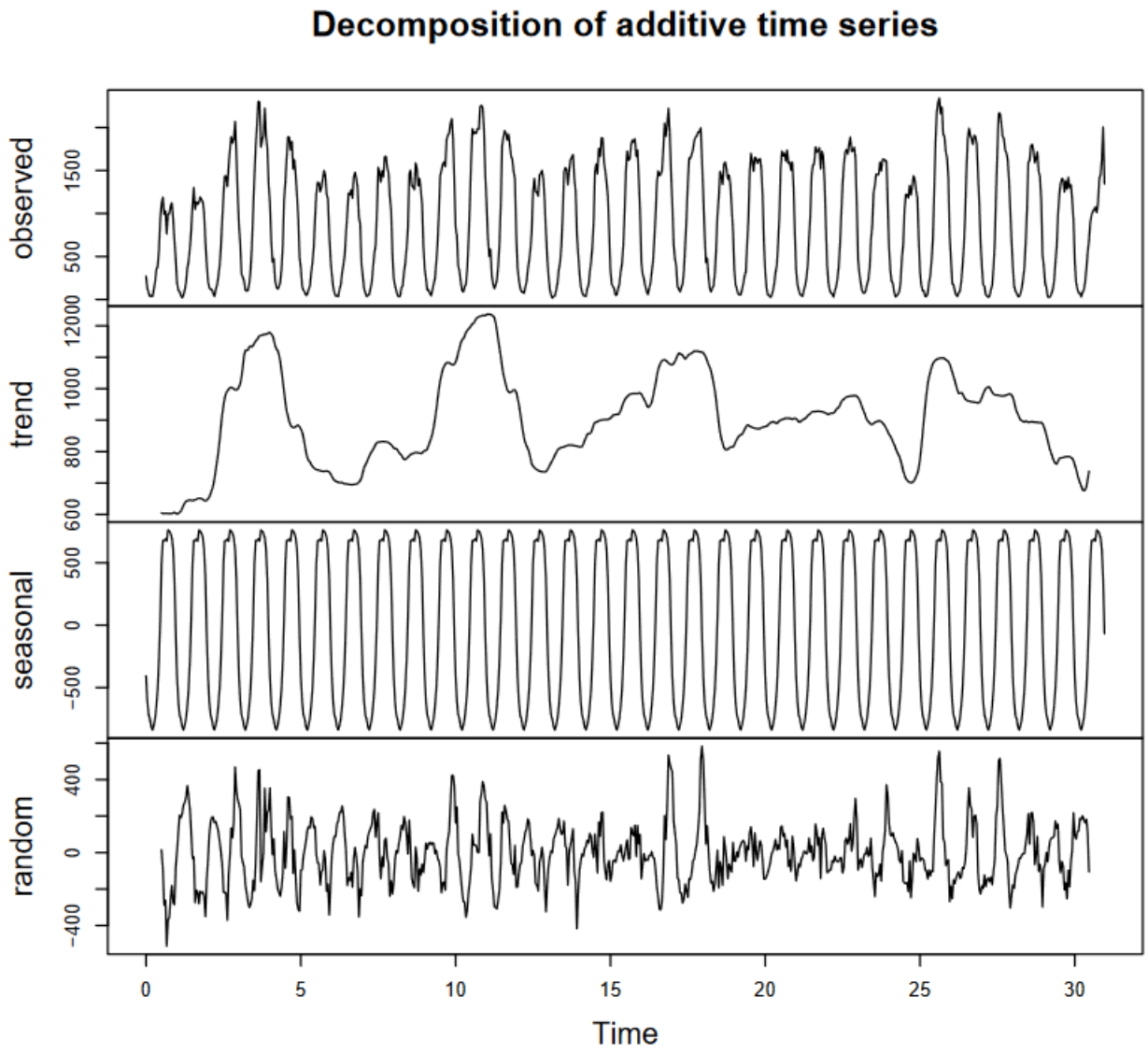
Explore the distribution of the hourly pedestrian count at each location using boxplots. Explore ways to make the resulting plot labels more readable.

```
> #set plot margins using plot parameter function  par()
> par(mar=c(18,4,2,2))
> #create box plots with label rotation in the x axis
> boxplot(December[3:83], las =2)
```



Plot a time series decomposition of pedestrian counts at the location “Melbourne.Central”.

```
> MC_TS = ts(December$Melbourne.Central, frequency = 24, start = c(0,1))
> decomp = decompose(MC_TS)
> plot(decomp)
```



Note, much of the data for the following questions has been sourced from http://www.statsci.org/datasets.html and links within.

1. Using the data sets provided as csv files and the lecture notes, try and reproduce all of the statistics and graphics from Lecture 1.

   Files are: {InvestA, InvestB, Toothbrush, Workers, Concrete time series}.csv

2. The following data records the length of rivers in the South Island of New Zealand. The lengths are given in kilometres. Data is grouped depending on where it flows into. Source: http://www.statsci.org/data/oz/nzrivers.html

Pacific Ocean:
209, 48, 169, 138, 64, 97, 161, 95, 145, 90, 121, 80, 56, 64, 209, 64, 72, 288, 322.

Tasman Sea:
76, 64, 68, 64, 37, 32, 32, 51, 56, 40, 64, 56, 80, 121, 177, 56, 80, 35, 72, 72, 108, 48.

(a) Calculate the summary stats for each group of rivers. Draw a boxplot.
(b) Test the hypothesis that rivers flowing into the Tasman Sea are shorter on average than those flowing into the Pacific Ocean. Use a significance of 1%

```
> PacificOcean <- c(209, 48, 169, 138, 64, 97, 161, 95, 145, 90, 121, 80, 56,
      64, 209, 64, 72, 288, 322)
> TasmanSea <- c(76, 64, 68, 64, 37, 32, 32, 51, 56, 40, 64, 56, 80, 121, 177,
      56, 80, 35, 72, 72, 108, 48)
>
> summary(PacificOcean)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   48.0    68.0    97.0   131.2   165.0   322.0

> summary(TasmanSea)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.00   48.75   64.00   67.68   75.00  177.00

> t.test(PacificOcean, TasmanSea, "greater", conf.level = 0.99)

      Welch Two Sample t-test

data:  PacificOcean and TasmanSea
t = 3.2632, df = 23.477, p-value = 0.001679
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 14.9219     Inf
sample estimates:
mean of x mean of y
131.15789  67.68182
```

3. When anthropologists analyze human skeletal remains, an important piece of information is living stature. Since skeletons are commonly based on statistical methods that utilize measurements on small bones, the following data was presented in a paper in the American Journal of Physical Anthropology to validate one such method. Variables are: MetaCarp – Metacarpal bone length in mm, Stature (Height of person) in cm. Source: http://www.statsci.org/data/general/stature.html

MetaCarp	Stature
45	171
51	178
39	157
41	163
48	172
49	183
46	173
43	175
47	173

Draw a scatterplot of the data with Stature as the vertical axis. Calculate the regression equation predicting Stature from MetaCarp. Comment on the accuracy of the model. Superimpose the line of best fit on your scatterplot.

```
> Metacarp <- c(45, 51, 39, 41, 48, 49, 46, 43, 47)
> Stature <- c(171, 178, 157, 163, 172, 183, 173, 175, 173)
> fitted = lm(Stature ~ Metacarp)
> plot(Metacarp, Stature) # scatterplot
> abline(fitted) # overlays the regression line
> summary(fitted) # gives a report on the regression fit

Call:
lm(formula = Stature ~ Metacarp)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0102 -3.1091 -1.1128  0.3891  7.4880

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   94.428     17.691   5.338  0.00108 **
Metacarp       1.700      0.388   4.380  0.00323 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.255 on 7 degrees of freedom
Multiple R-squared:  0.7327,  Adjusted R-squared:  0.6945
F-statistic: 19.19 on 1 and 7 DF,  p-value: 0.003234
```

4. The ocean swell produces spectacular eruptions of water through a hole in the cliff at Kiama, about 120km south of Sydney, known as the Blowhole. The times at which 65 successive eruptions occurred from 1340 hours on 12 July 1998 were observed using a digital watch. Source: http://www.statsci.org/data/oz/kiama.html

   Challenge: download the data into R directly from: http://www.statsci.org/data/oz/kiama.txt (See ATHR page 18) or alternatively use the file: Data: kiama.txt

   Read these data into R, creating a vector named 'kiama'. Calculate the mean, standard deviation. Draw the default histogram. Using help, try and draw an improved histogram of your own design by changing range, class intervals and colour etc.

```
> kiama <- read.table("http://www.statsci.org/data/oz/kiama.txt", header=TRUE)
> summary(kiama)
    Interval
 Min.   :  7.00
 1st Qu.: 14.75
 Median : 28.00
 Mean   : 39.83
 3rd Qu.: 60.00
 Max.   :169.00
```

5. The timber data are for specimens of 50 varieties of timber, for modulus of rigidity, modulus of elasticity and air dried density, arranged in increasing order of magnitude of the density. Source: http://www.statsci.org/data/oz/timber.html

   Read these data into R, creating a data frame named 'timber'. You can use the data file: timber.txt or load directly from: http://www.statsci.org/data/oz/timber.txt

   (a) which variable: elasticity or density is a better predictor of rigidity?

```
> timber <- read.delim("H:/timber.txt")
>   View(timber)
> cor(timber)
          Rigid     Elast      Dens
Rigid 1.0000000 0.8183311 0.8587485
Elast 0.8183311 1.0000000 0.7388302
Dens  0.8587485 0.7388302 1.0000000
```

   (b) using your choice of variable calculate the regression equation predicting rigidity, draw a scatterplot of the data, showing the fitted model.fitt

```
> fitted = lm(timber$Rigid ~ timber$Dens)
> summary(fitted)

Call:
lm(formula = timber$Rigid ~ timber$Dens)

Residuals:
    Min      1Q  Median      3Q     Max
-520.72  -96.04    5.50  100.14  599.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  206.723    129.866   1.592    0.118
timber$Dens   30.050      2.588  11.611 1.53e-15 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 217.3 on 48 degrees of freedom
Multiple R-squared:  0.7374,  Adjusted R-squared:  0.732
F-statistic: 134.8 on 1 and 48 DF,  p-value: 1.525e-15
```

(c) challenge: calculate the regression equation predicting rigidity as a function of both elasticity and density. Comment on the quality of your model vs the single predictor in (b).

```
> fitted = lm(timber$Rigid ~ timber$Dens + timber$Elast)
> summary(fitted)

Call:
lm(formula = timber$Rigid ~ timber$Dens + timber$Elast)

Residuals:
    Min      1Q  Median      3Q     Max
-451.54  -81.46  -22.87   67.39  776.33

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.8300   121.1577  -0.015    0.988
timber$Dens   19.5830     3.2851   5.961 3.08e-07 ***
timber$Elast   3.4179     0.7925   4.313 8.21e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 185.9 on 47 degrees of freedom
Multiple R-squared:  0.8119,  Adjusted R-squared:  0.8039
F-statistic: 101.4 on 2 and 47 DF,  p-value: < 2.2e-16

Note that fitting rigidity against both elasticity and density gives a higher
      degree of fit: R-squared ~ 80%, single predictor (density) ~ 73%
```

6. Challenge: Using the data: InvestA.csv draw a boxplot. You will need to use the help file to work out the syntax – try ?boxplot as a starting point...

```
> boxplot(InvestA$FV ~ InvestA$Group)
```

Using the data: InvestA.csv, now use the 'aggregate' function to calculate the mean of each group. This is similar to the 'tapply' function but returns a data frame. Use help to work out the syntax...

```
> aggregate(InvestA$FV, by = list(InvestA$Group), FUN = "mean")
  Group.1         x
1       1  689.3454
2       2  874.0045
3       3  802.4651
4       4 1339.0980
5       5  786.7376
6       6  797.1628
```

7. Analyse Victorian Retail Turnover: for Food retailing; Household goods retailing; Clothing, footwear and personal accessory retailing; Department stores; Other retailing; Cafes, restaurants and takeaway food service for the period Jan 2010 – Dec 2020 using Australian Bureau of Statistics data. You will need to copy the data from the Excel file: 8501.0 Retail Trade, Australia.xls.

   Draw a time series plot of the data and plot the time series decomposition. Comment on the main elements in the time series.

   Can you see any COVID-19 effects during 2020?

```
#### One Approach
#### First copy and paste the Data1 page of the 8501.0 Retail Trade,
      Australia.xls file.
#### Save it as a .csv file, absdata.csv.
#### Then load the .csv file TO R.

df = read.csv("absdata.csv")

# Clothing
series  = ts(as.numeric(df[343:474,4]), frequency = 12, start = c(2010,1))
plot(series)
decomposed_series = decompose(series)
plot(decomposed_series)

# Department Stores
series  = ts(as.numeric(df[343:474,5]), frequency = 12, start = c(2010,1))
plot(series)
decomposed_series = decompose(series)
plot(decomposed_series)

# Cafe Restaurant & Takeaway food services
series  = ts(as.numeric(df[343:474,7]), frequency = 12, start = c(2010,1))
plot(series)
decomposed_series = decompose(series)
plot(decomposed_series)

# By looking the data we can see that Cafe Restaurant & Takeaway food services
# are affected by COVID-19 more than the other industries.
# We can see the change in the direction of the trend and also unexpected drop
      in the observed values...
```