2020 Project

Thinking About Data

Ryan G - 17805315

Declaration

- I hold a copy of this assignment that we can produce if the original is lost or damaged.
- I hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
- No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.
- I am aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (which may retain a copy on its database for future plagiarism checking).
- I hereby certify that we have read and understand what the School of Computing and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

Problem

> The health minister of Australia has been concerned about the increase in new cases of people infected with the Tiara virus. They have hired you as a consultant to examine their data on the new cases of Tiara virus and the flu. The data is provided to you in a CSV file called project2020A.csv containing observations of the daily new cases of both the Tiara and flu viruses for each state in Australia over a 100 day period. The variables are:
>
> - `city`: The city in which the new cases of the virus is counted.
> - `newTiara`: The number of new cases of the Tiara virus for the day.
> - `newFlu`: The number of new cases of the flu virus for the day.
> - `date`: the day number starting from 1 and ending at 100.
>
> The health minister wants statistics measured from the data so that they can be reported to the public. The five required pieces of analysis are presented below. # Preamble Before beginning it is necessary to set the working directory, load any necessary packages and load the data set.

```
## Preamble
# setwd("~/Dropbox/Notes/DataSci/ThinkingAboutData/Assessment/")
## Install Pacman
load.pac <- function() {

  if(require("pacman")){
    library(pacman)
  }else{
    install.packages("pacman")
    library(pacman)
  }
```

```
    pacman::p_load(xts, sp, gstat, ggplot2, rmarkdown, reshape2, ggmap,
                   parallel, dplyr, plotly, tidyverse, reticulate, UsingR, Rmpfr,
                   swirl, corrplot, gridExtra, mise, latex2exp, tree, rpart, lattice,
                   coin, primes, epitools, maps, clipr, ggmap, RColorBrewer)


   mise()
   select <- dplyr::select
}

load.pac()
```

```
## Loading required package: pacman
```

```
setwd(dir = "/home/ryan/Notes/DataSci/ThinkingAboutData/")
load(file = "~/Notes/DataSci/ThinkingAboutData/TAD.rdata")
load(file = "./TAD.rdata")
print("Success")
```

```
## [1] "Success"
```

The data can be inspected thusly:

```
(read.csv("../0datasets/project2020A.csv") -> data)  %>% head()
```

```
##     city newTiara newFlu date
## 1 Sydney        1     40    1
## 2 Sydney        2     43    2
## 3 Sydney        4     35    3
## 4 Sydney        0     38    4
## 5 Sydney        3     37    5
## 6 Sydney        4     31    6
```

```
str(data)
```

```
## 'data.frame':    400 obs. of  4 variables:
##  $ city    : chr  "Sydney" "Sydney" "Sydney" "Sydney" ...
##  $ newTiara: int  1 2 4 0 3 4 2 3 3 3 ...
##  $ newFlu  : int  40 43 35 38 37 31 43 38 49 35 ...
##  $ date    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
summary(data)
```

```
##      city              newTiara           newFlu           date
##  Length:400         Min.   :    0.00   Min.   : 9.00   Min.   :  1.00
##  Class :character   1st Qu.:   31.75   1st Qu.:21.75   1st Qu.: 25.75
##  Mode  :character   Median :  389.00   Median :29.00   Median : 50.50
##                     Mean   : 2623.00   Mean   :29.71   Mean   : 50.50
##                     3rd Qu.: 3130.25   3rd Qu.:37.00   3rd Qu.: 75.25
##                     Max.   :29711.00   Max.   :57.00   Max.   :100.00
```

```
if(sum(is.na(data)) > 0) {
  print("The data Needs to be Cleaned")
} else {
  print("The data does not require cleaning")
}
```

```
## [1] "The data does not require cleaning"
```

This data set has provides 400 observations with 4 features, one of which is categorical. There is no missing data in this data set.

Question 1

> Assume that the number of new flu cases each day are independent over the set of days. Test if the mean number of new flu cases over the set of days is different for each city, and if so determine which cities have a statistically different mean.

It is first necessary to get the data into a tidy format, so first encode the categorical variable as a `factor`:

```
data$city  <- factor(data$city)
```

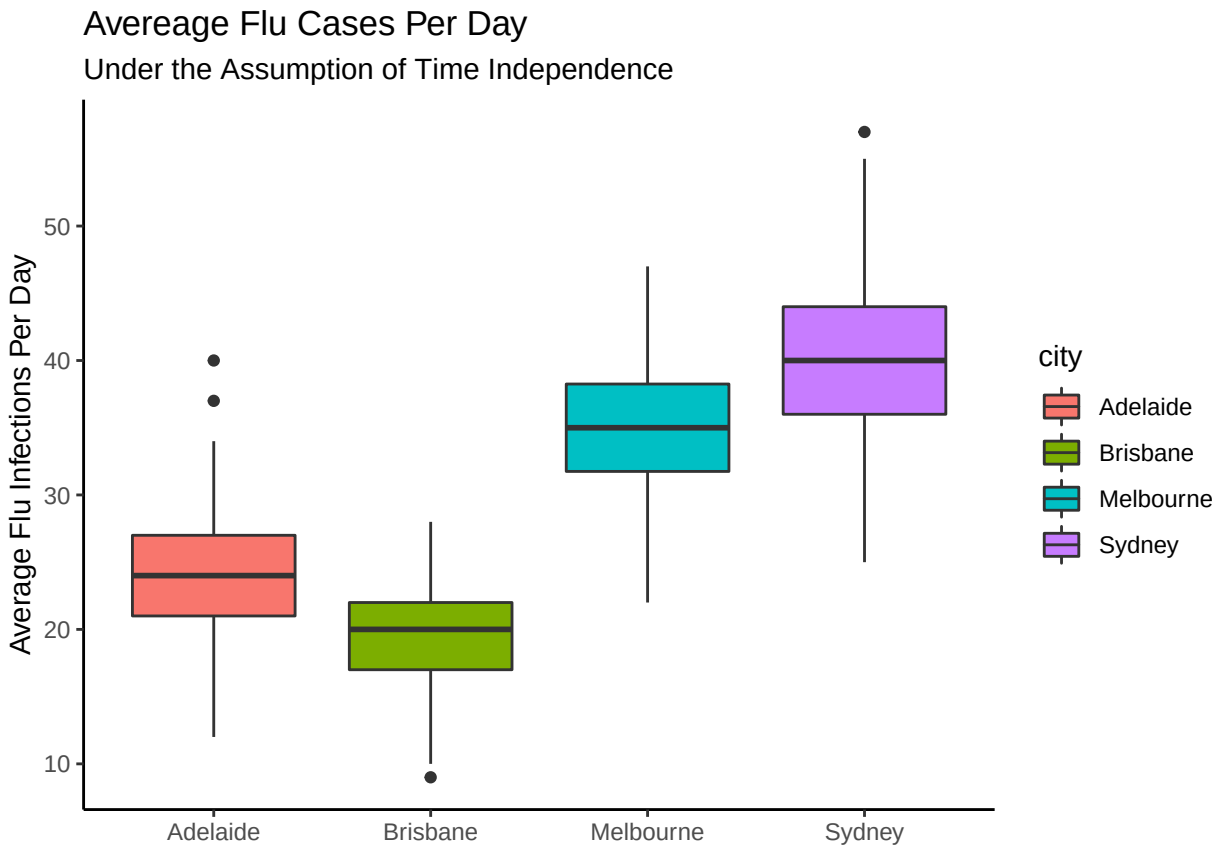
Now aggregate the data:

```
mean_city <- aggregate(newFlu ~ city, data, mean, na.rm = TRUE)
```

Plot

This aggregated data can be plotted, illustrating the average number of new flu infections accross cities under the assumption that the rate of infection is independent of time.

```
p <- ggplot(data, aes(x = city, y = newFlu, fill = city)) +
  theme_classic() +
  labs(title = "Avereage Flu Cases Per Day",
       subtitle = "Under the Assumption of Time Independence",
y = "Average Flu Infections Per Day") +
  theme(axis.title.x = element_blank())

p + geom_boxplot()
```



Observations from Plot

This strongly suggests that the number of infections in Sydney and Melbourne are higher than in Adelaide or Brisbane, it would be reasonable to expect that Sydney would have a statistically higher mean value of new cases.

Analysis and Results

In order to assess whether or not the mean value does differ accross these cities a hypothesis test will be established.

Hypothesis

- H_0 : The mean value accross populations does not change
  - And hence we would expect the mean value to be the overall mean
- H_a : There is a difference between the mean values across cities.

Test statistic

The F statistic is given by equation (1) and compares the variance within groups to the variance outside groups:

$$F = \frac{\mathrm{SS}_B/(K-1)}{\mathrm{SS}_W(K-1)} \tag{1}$$

$$SS_B = \sum_{i=1}^{K} n_k \left(\bar{x}_k - \bar{x}\right)^2 \tag{2}$$

$$SS_W = \sum_{i=1}^{K} \left(n_k - 1\right) s_k^2 \tag{3}$$

where:

- k is the group number or city.
- K is the number of groups (In this case 4 cities)
- SS_B is the sum of squared differences from the group means to the overall means as defined in equation (2)
- SS_W is the sum of squared differences between group values and group mean as defined in equation (3)

The F statistic can also be calculated in ***R*** using the `oneway` function like so:

```
(F_obs <- oneway.test(newFlu ~ city, data, var.equal = FALSE))
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  newFlu and city
## F = 369.47, num df = 3.00, denom df = 217.21, p-value < 2.2e-16
```

```
F_obs <- F_obs$statistic
```

Rejection Region

Rather than using the F statistic directly, the statistic of concern will be the probability of a Type I error (α) which is essentially a false positive.

The null hypothesis will be rejected for an α value less than 5%, this represents a low probability of a type I error which is good evidence for rejecting the null hypothesis.

This value was reported above by the `oneway.test` function but it will be derived from first principles for want of rigour.

Statistic

The p-value is the measured probability of a type I error, it can be measured by simulating the data under the assumption that the null hypothesis is true and measuring the frequency at which the null hypothesis would be rejected by mere chance, that frequency will be accepted as the probability of a type I error.

In order to simulate the data, the observations can be permuted in order remove any meaningful difference between mean values that would violate the null hypothesis and the F statistic measured. The frequency at which a more extreme F value is observed is the p value, this is shown below:

```
x <- replicate(10^3, {
  ## Permute the Categories to satisfy H_0
  city_perm <- sample(data$city)
  ## Calculate the F-Statistic
#  F_sim <- oneway.test(newFlu ~ city, data, var.equal = FALSE)$statistic

  ## Calculate Summary Statistics
  K <- length(unique(data$city))
  sd_within_groups <- aggregate(newFlu ~ city, data, sd)$newFlu^2
  xbar <- mean(data$newFlu)
  xbar_within_groups <- aggregate(newFlu ~ city, data, mean)$newFlu

  ## Calculate Squared Sums
  SSB <- length(xbar_within_groups)*(xbar_within_groups-xbar)^2
  SSW <- sum((length(sd_within_groups)-1)*sd_within_groups)

  ## Divide to get F
  F_sim <- (SSB/(K-1) ) / (SSW/(K-1))

  ## Is this more extreme than what we saw?
  F_sim > F_obs
  })

## Average the values
mean(x)
```

```
## [1] 0
```

This returns a p -value of 0 which is consistent with the built in output of the `oneway` function.

Conclusion

The probability of rejecting the null hypothesis is very small, this is good evidence to support rejecting the null hypothesis, and because the p value is smaller than the threshold we accept the alternative hypothesis.

This probability does not provide us sufficient information however, to determine the probability of correctly accepting the null hypothesis $(1 - \beta)$.

Question 2

> After more investigation, it was found that the sample data was collected from the set of people who visitedthe major city hospital in the last year. The number of people involved in the study is provided below.Test if there is a difference in proportions of the total number of new cases (over the 100 days) between Melbourne and Sydney.

	Participants
Sydney	40, 000
Melbourne	35, 000
Brisbane	20, 000
Adelaide	25, 000

Because it's just sydney and melbourne we can do a t test

Plot

The proportion of new cases in Sydney and Melbourne may be determined by selecting the correct variables from the data, filtering out observations from Sydney and Melbourne and then dividing by the number of participants:

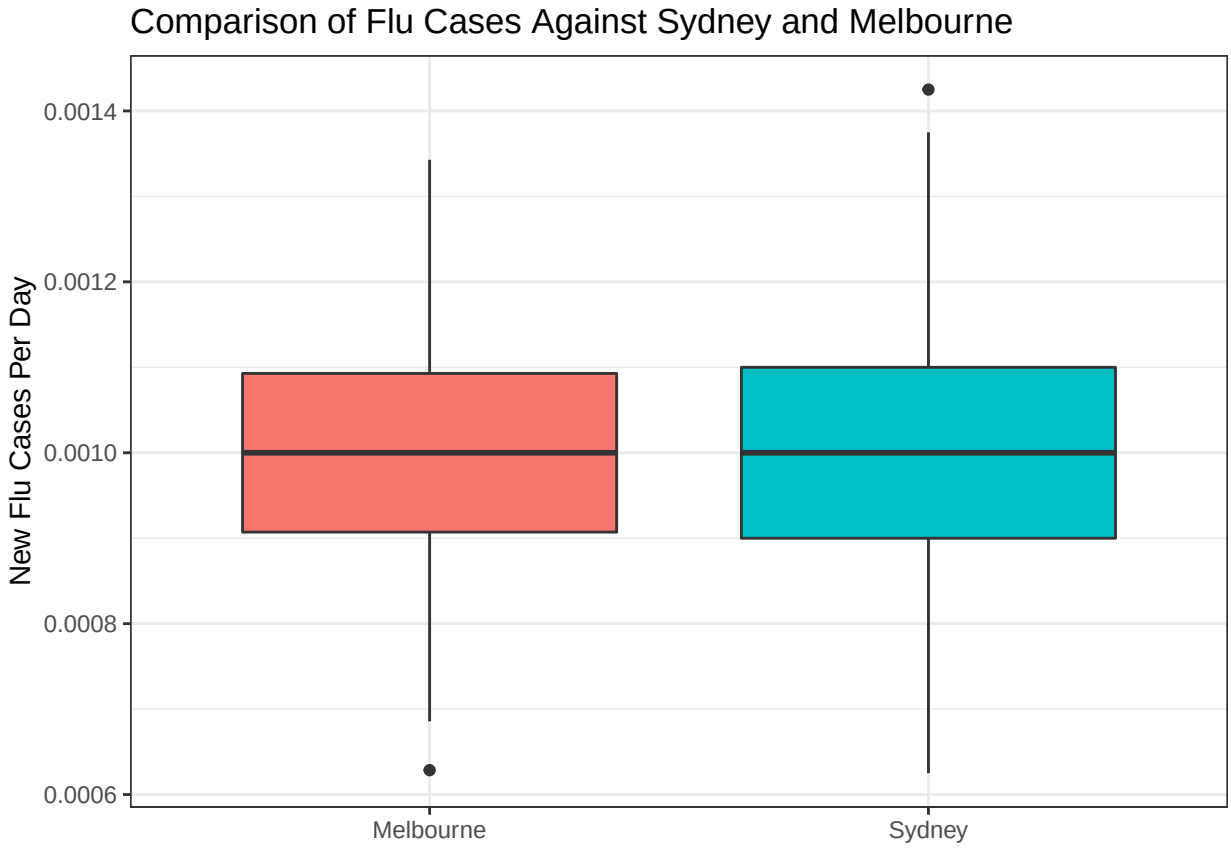
```
select <- dplyr::select
filter <- dplyr::filter

tb <- table(data$city)

prop_df <-data %>%
 filter(city %in% c("Sydney", "Melbourne")) %>%
 select(city, newFlu) %>%
 mutate(newFlu = newFlu/c(rep(40000, tb["Sydney"]),
                          rep(35000, tb["Melbourne"]))
        )

aggregate(newFlu ~ city, prop_df, mean)
```

```
##        city      newFlu
## 1 Melbourne 0.001000286
## 2    Sydney 0.001008250
```

This provies that the number of new infections per day is at a rate of approximately 0.1%.

From this a boxplot can be produced to compare the two proportions:

```
ggplot(prop_df, aes(x = city, y = newFlu, fill = city)) +
  geom_boxplot() +
  theme_bw() +
    theme(axis.title.x = element_blank()) +
  guides(fill = FALSE) +
  labs(y = "New Flu Cases Per Day", title = "Comparison of Flu Cases Against Sydney and Melbourne")
```



Observations from Plot

The plot does not suggest that there is any difference between the proportion of cases between Sydney and Melbourne.

Analysis and Results

Student's t-distribution

The *Central Limit Theorem* provides that the distribution of mean values from sample of a population will be normally distributed such that $\overline{X} \sim \mathcal{N}\left(0, \frac{s}{\sqrt{n}}\right)$ if:

- those samples are sufficiently large, or
- the population is normally distributed

This means that a standardised value for the distribution of mean values can be used to measure the p-value as shown in equation (5).

can be compared using a *Student's t* test)

$$z_i = \frac{x_i - \overline{x}}{s} \tag{4}$$

$$\implies t = \frac{(\overline{x_1} - \overline{x_2}) - 0}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{5}$$

this is built into ***R*** and can be implemented with the `t.test` function:

```
t.test(newFlu ~ city, prop_df)
```

```
##
##  Welch Two Sample t-test
##
## data:  newFlu by city
## t = -0.36942, df = 197.98, p-value = 0.7122
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.047898e-05  3.455041e-05
## sample estimates:
## mean in group Melbourne    mean in group Sydney
##             0.001000286             0.001008250
```

This provides a large p-value indicating a high probability that any differences in the sample observations are a result of mere chance rather than indicative of a difference in population means.

Simulation

This result can also be simulated in order to

```
set.seed(85284)

mean_diff_obs <- aggregate(newFlu ~ city, prop_df, mean)[,2] %>% diff()

xbar_sim <- replicate(10^3, {
  city_perm <- sample(prop_df$city)
  mean_diff_sim <- aggregate(newFlu ~ city_perm, prop_df, mean)[,2] %>% diff()

  # Is this more extreme? Is it a false pos?
  abs(mean_diff_sim) > abs(mean_diff_obs)
})

# What Proportion are false postive?
mean(xbar_sim)
```

```
## [1] 0.719
```

This shows, assuming there is no difference between the two populations, that the probability of detecting such a change is $\approx 72\%$, this is consistent with the t-test from before.

Conclusion

A p-value in excess of 0.7 is very large and indicates that there is insufficient evidence to reject the hypothesis that there is a difference between the mean value of the proportion of new infections between cities.

Hence it is *not* concluded that there is any difference.

Question 3

Plot

Observations from Plot

Analysis and Results

Conclusion

Question 4

Plot

Observations from Plot

Analysis and Results

Conclusion

Question 5

Plot

Observations from Plot

Analysis and Results

Conclusion