Assignment 1

Ben Lu

2024-07-24

Question 1

As there are three data sets for the annual cyclists count data, there may be human errors when they entered the name of the streets. The name of the streets may have slightly different variations such as “Curran Street” or “Curran St”, and we would prefer the same streets to have the same name “Curran Street”.

Hence, the main challenge to tidying our data would be finding streets which are the same and combining them into one. This is challenging because it may be difficult to combine some of the street name variations into one, further manual editing may be required.

Question 2

```
#read csv files
cycle_2016_data <- read_csv("dailyakldcyclecountdata2016_updated.csv", show_col_types = FALSE) %>%
  mutate(Date = dmy(Date))

cycle_2017_data <- read_csv("dailyakldcyclecountdata2017_1.csv", show_col_types = FALSE) %>%
  mutate(Date = dmy(Date))

cycle_2018_data <- read_csv("dailyakldcyclecountdata2018.csv", show_col_types = FALSE) %>%
  mutate(Date = dmy(Date))

rain_2016_2017_data <- read_csv("rain2016-17.txt", skip = 9, show_col_types = FALSE)
rain_2018_data <- read_csv("rain2018.txt", skip = 9, show_col_types = FALSE)
rain_data <- rbind(rain_2016_2017_data, rain_2018_data) %>%
  rename("Date" = "Date(NZST)") %>%
  mutate(Date = ymd(Date))

#generate summaries for total cyclists
cycle_count_2016 <- cycle_2016_data %>%
  summarise(
    Date,
    Total_Cyclists = rowSums(cycle_2016_data[, -1], na.rm = TRUE))

cycle_count_2017 <- cycle_2017_data %>%
  summarise(
    Date,
    Total_Cyclists = rowSums(cycle_2017_data[, -1], na.rm = TRUE))

cycle_count_2018 <- cycle_2018_data %>%
  summarise(
    Date,
    Total_Cyclists = rowSums(cycle_2018_data[, -1], na.rm = TRUE))

cycle_summary <- rbind(cycle_count_2016, cycle_count_2017, cycle_count_2018)

#generate summaries for total rainfall
#calculate the average rainfall for each day in Mangere
rain_mangere <- rain_data %>%
  filter(Station == "22719") %>%
  group_by(Date) %>%
  summarize(rain_mangere_mean = mean(`Amount(mm)`, na.rm = TRUE))

#calculate the average rainfall for each day in North Shore Albany
rain_albany <- rain_data %>%
  filter(Station == "37852") %>%
  group_by(Date) %>%
  summarize(rain_albany_mean = mean(`Amount(mm)`, na.rm = TRUE))

#combine data frame and calculate average rainfall of Mangere and Albany
rain_summary <- rain_mangere %>%
  inner_join(rain_albany, by = "Date")

print(slice(rain_summary, 1:10))
```

```
## # A tibble: 10 × 3
##    Date       rain_mangere_mean rain_albany_mean
##    <date>                 <dbl>            <dbl>
##  1 2016-01-01           0.612             1.08
##  2 2016-01-02           0.721             0.875
##  3 2016-01-03           0.338             0.229
##  4 2016-01-04           0.00417           0
##  5 2016-01-05           0                 0
##  6 2016-01-06           0                 0
##  7 2016-01-07           0                 0
##  8 2016-01-08           1.50              1.54
##  9 2016-01-09           0.00833           0
## 10 2016-01-10           0                 0
```

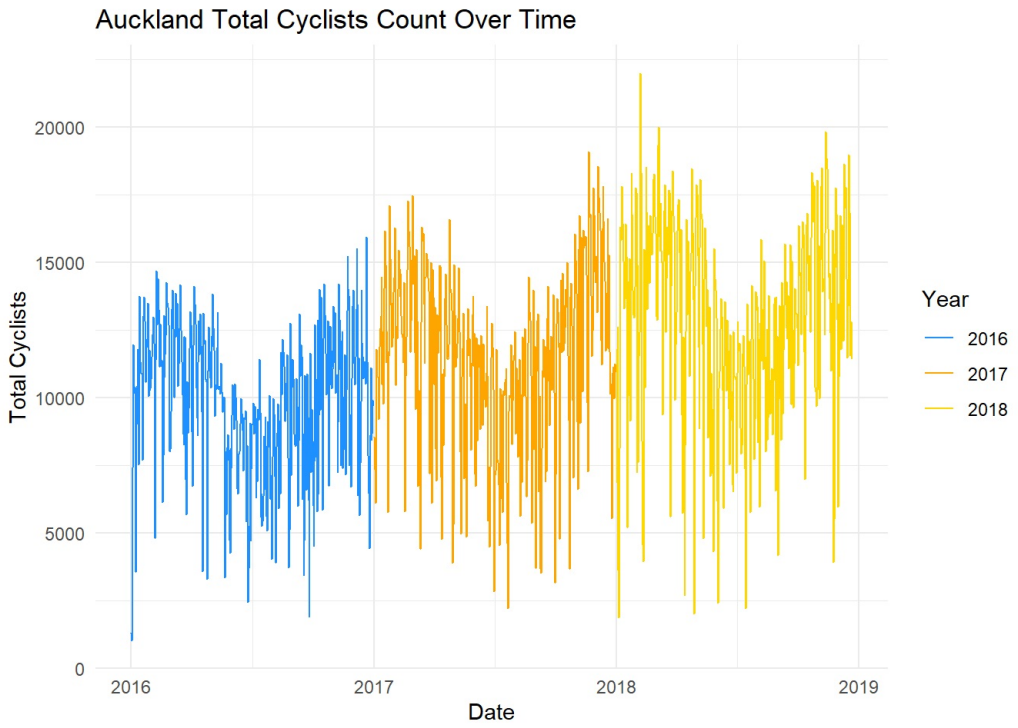
```
rain_summary <- rain_summary %>%
  summarise(
     Date = Date,
    `Mean_Rain(mm)` = rowMeans(rain_summary[,2:3]))

#combine summaries into one data frame, remove na rows
summary_data <- cycle_summary %>%
  inner_join(rain_summary, by = "Date") %>%
  drop_na()

#process data frame, create month, weekday and season columns
summary_data <- summary_data %>%
  mutate(Date = as.Date(Date),
         Month = month(Date, label = TRUE, abbr = FALSE),
         Weekday = wday(Date, label = TRUE),
         Year = as.factor(year(Date)),
         Season = case_when(
           Month %in% c("December", "January", "February") ~ "Summer",
           Month %in% c("March", "April", "May") ~ "Autumn",
           Month %in% c("June", "July", "August") ~ "Winter",
           Month %in% c("September", "October", "November") ~ "Spring"),
         Weekday = factor(Weekday, levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"), order = FALSE),
         Season = factor(Season, levels = c("Summer", "Autumn", "Winter", "Spring")))

print(slice(summary_data, 1:10))
```

```
## # A tibble: 10 × 7
##    Date       Total_Cyclists `Mean_Rain(mm)` Month   Weekday Year  Season
##    <date>              <dbl>           <dbl> <ord>   <fct>   <fct> <fct>
##  1 2016-01-01           1299         0.844   January Fri     2016  Summer
##  2 2016-01-02           1030         0.798   January Sat     2016  Summer
##  3 2016-01-03           7423         0.283   January Sun     2016  Summer
##  4 2016-01-04          11956         0.00208 January Mon     2016  Summer
##  5 2016-01-05          10167         0       January Tue     2016  Summer
##  6 2016-01-06          10387         0       January Wed     2016  Summer
##  7 2016-01-07           9573         0       January Thu     2016  Summer
##  8 2016-01-08           3535         1.52    January Fri     2016  Summer
##  9 2016-01-09           8998         0.00417 January Sat     2016  Summer
## 10 2016-01-10          10429         0       January Sun     2016  Summer
```

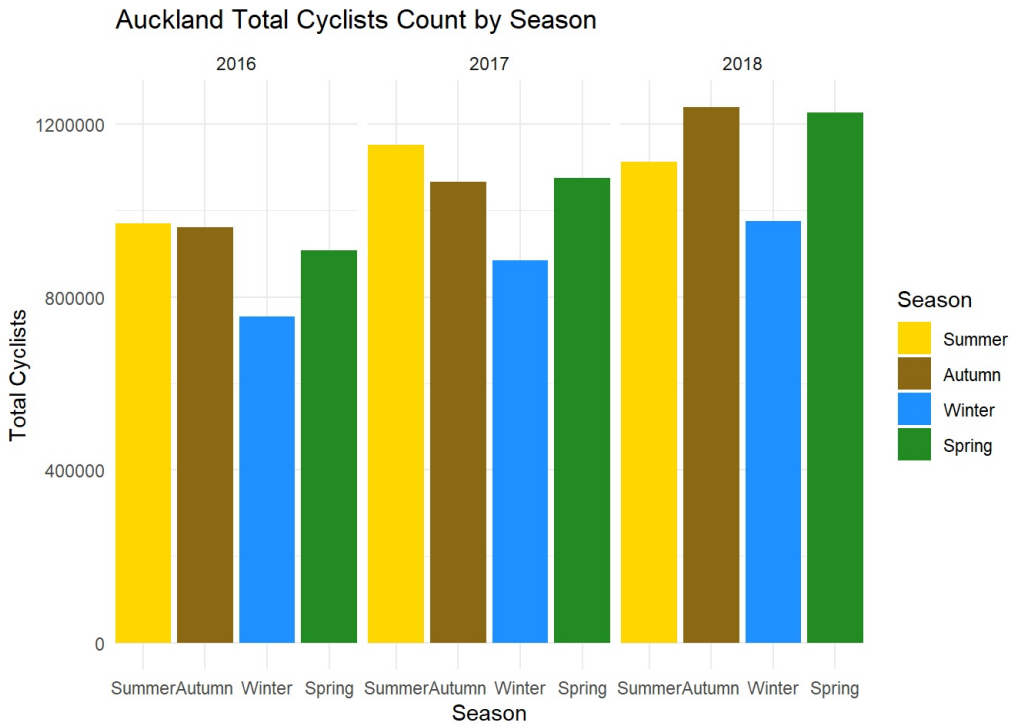
As the rain could be scattered across different stations in Auckland, it could be raining for a few hours in Albany and a few hours in Mangere, we prefer not to take the daily sum of rainfall because the amount of rain is inconsistent across different stations. Hence, we prefer to take the daily average of the rainfall in Mangere and Albany, and then take the average of the rainfall in both locations to give us the average daily rainfall.

Question 3

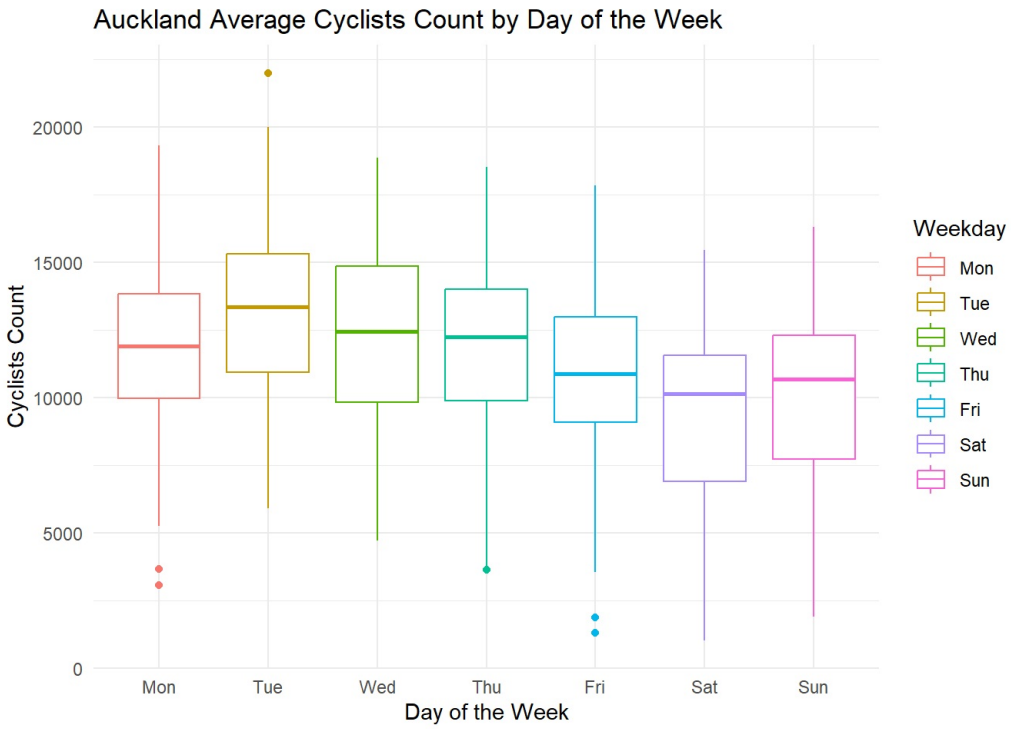
```
#Cyclists over time
ggplot(summary_data, aes(x = Date, y = Total_Cyclists, color = Year)) +
  geom_line() +
  scale_color_manual(values = c("2016" = "dodgerblue", "2017" = "orange", "2018" = "gold")) +
  labs(title = "Auckland Total Cyclists Count Over Time", x = "Date", y = "Total Cyclists") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14)) +
  theme_minimal()
```




```
#cyclists by season
ggplot(summary_data, aes(x = Season, y = Total_Cyclists, fill = Season)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("Summer" = "gold", "Spring" = "forestgreen", "Winter" = "dodgerblue", "Autumn" = "
goldenrod4")) +
  facet_wrap(~Year, ncol = 3) +
  labs(title = "Auckland Total Cyclists Count by Season", x = "Season", y = "Total Cyclists") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
        panel.spacing.x = unit(20, "lines"),
        axis.text.x = element_text(margin = margin(t = 10))) +
  scale_x_discrete(expand = expansion(mult = c(0.1, 0.1))) +
  theme_minimal()
```

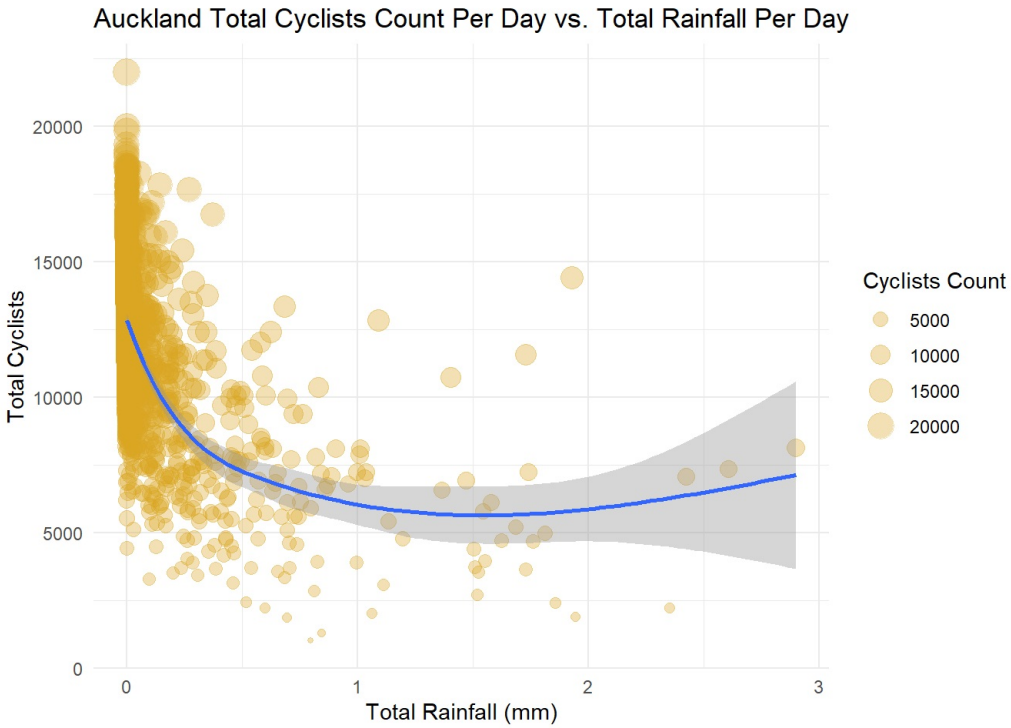



```
#cyclists by weekday
ggplot(summary_data, aes(x = Weekday, y = Total_Cyclists, colour = Weekday)) +
  geom_boxplot() +
  labs(title = "Auckland Average Cyclists Count by Day of the Week", x = "Day of the Week", y = "Cyclists Count")
+
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14)) +
  theme_minimal()
```




```
#cyclists by rain
ggplot(summary_data, aes(x = `Mean_Rain(mm)`, y = Total_Cyclists)) +
  geom_point(aes(size = Total_Cyclists), alpha = 1/3, colour = "goldenrod") +
  geom_smooth() +
  labs(title = "Auckland Total Cyclists Count Per Day vs. Total Rainfall Per Day", x = "Total Rainfall (mm)", y =
"Total Cyclists", size = "Cyclists Count") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14)) +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Question 4

```
#fit model
model_fit <- lm(Total_Cyclists ~ Year + Season + Weekday + `Mean_Rain(mm)`, data = summary_data)
summary(model_fit)
```

```
##
## Call:
## lm(formula = Total_Cyclists ~ Year + Season + Weekday + `Mean_Rain(mm)`,
##     data = summary_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8534.3 -1232.4   200.3  1348.0  9475.3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11763.0      225.5  52.165  < 2e-16 ***
## Year2017          1719.4      158.1  10.879  < 2e-16 ***
## Year2018          3175.0      159.2  19.945  < 2e-16 ***
## SeasonAutumn      -394.1      184.5  -2.137 0.032853 *
## SeasonWinter     -2755.2      184.4 -14.938  < 2e-16 ***
## SeasonSpring      -791.6      184.8  -4.282 2.01e-05 ***
## WeekdayTue        1454.7      242.7   5.995 2.78e-09 ***
## WeekdayWed         916.3      243.0   3.770 0.000172 ***
## WeekdayThu         483.7      242.9   1.992 0.046673 *
## WeekdayFri        -716.7      242.3  -2.958 0.003164 **
## WeekdaySat       -2251.9      242.4  -9.291  < 2e-16 ***
## WeekdaySun       -1607.9      242.8  -6.623 5.54e-11 ***
## `Mean_Rain(mm)`  -5484.6      201.3 -27.239  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2136 on 1074 degrees of freedom
## Multiple R-squared:  0.6217, Adjusted R-squared:  0.6175
## F-statistic: 147.1 on 12 and 1074 DF,  p-value: < 2.2e-16
```

Question 5

There are a few factors which influences the number cyclists in Auckland. In observing our plot for the total number of cyclists over time, there appears to be seasonal fluctuations of peaks and troughs. This suggests that in some seasons the number of cyclists are higher, and lower for others. This is expected because the amount of rainfall vary between seasons. Furthermore, it also appears the number of cyclists also increases over time. This is supported by evidence from our model. As we obtained a very small p-value of 2e-16 for 2017 and 2018, it suggests that the yearly factors are statistically significant. In which case, the years 2017 and 2018 increases the number of cyclists by approximately 1719 and 3175 compared to 2016, respectively.

In observing our bar chart for the number of cyclists categorized by season, it clearly shows that winter has the lowest number of cyclists for every year. This is likely due to the fact that during winter, the chance of rain increases, and therefore the number of people cycling also decreases. In comparison, summer has the highest number of cyclists for 2016 and 2017, which is likely because during summer, the chance of rain decreases. However, for 2018, summer had a lower number of cyclists than spring and autumn which is unexpected. Regardless, there is clearly a seasonal effect. In our model, we obtained an extremely small p-value of 2e-16 for the season of winter, which indicates statistical significance. We estimate that for every winter, the number of cyclists decreases by approximately 2755 in comparison to summer. Similarly, for Autumn and Spring which are also statistically significant, they have an impact of a 394 and 792 decrease in the number of cyclists compared to summer, respectively. This is likely due to the presence of more rain in these seasons.

In observing our scatter-plot for the total number of cyclists and the average amount of rainfall (mm), most of the data is clustered on the left end where the average amount of rainfall is lower. The trend line indicates that the number of cyclists decreases dramatically as the average amount of rainfall increases to approximately 0.5mm, from which point on-wards, it begins to flatten and curve slightly upwards. This means that as the amount of rain increases to 0.5mm, there is a huge drop in the number of cyclists, however, from that point on-wards, there would still be certain groups who continue to cycle despite the increasing amount of rain. In our model, we obtained an extremely small p-value of 2e-16 which indicates that the average amount of rainfall is statistically significant. We estimate that for every mm increase in the average amount of rain, the number of cyclists decreases by approximately 5485. In conclusion, rainfall has a huge impact on the number of cyclists in Auckland.