Example Solution

Lauren Kennedy

25/08/2021

These are the packages you will need to complete this assignment

```
library(MASS)
library(ggplot2)
library(GGally)
library(tidyverse)
library(here)
library(viridis)
```

Read in your data

```
electricity <- read_csv(here("data/electricity_data.csv"))
view(electricity)

electricity_long <- pivot_longer(electricity, cols = starts_with("E_"),names_to = 'Variable', va
lues_to = 'Value')
view(electricity_long)

survey_data <- read_csv(here("data/survey_data.csv"))
survey_data <-survey_data %>%
  rename(respondent= Respondent_Serial)
view(survey_data)

important_dates <- read_csv(here("data/important_dates.csv"))
view(important_dates)

weather_data <- read_csv(here("data/maxtemp_data.csv"))
view(weather_data)
```

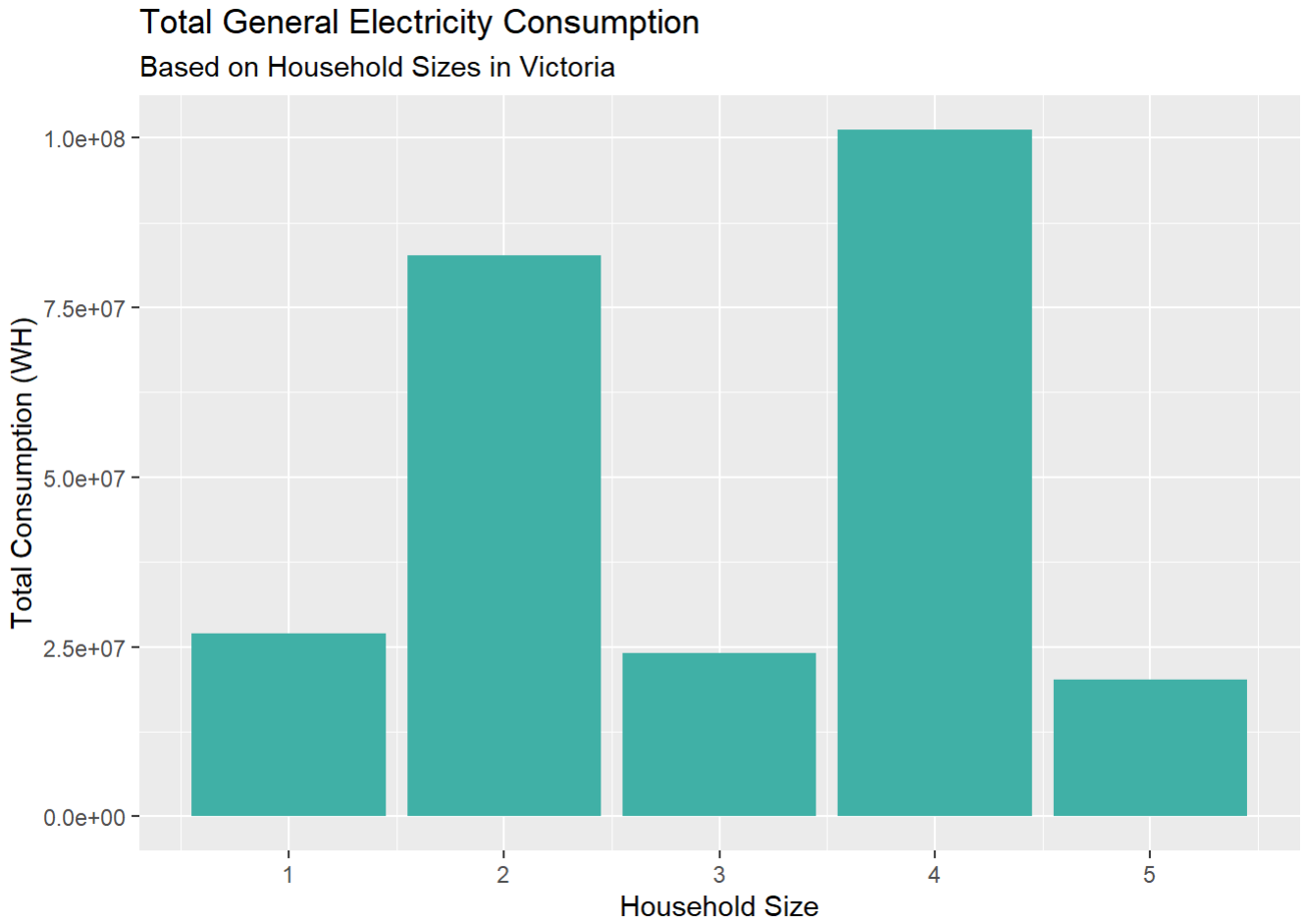
Figure 1

```
electricity_survey <- full_join(electricity_long,survey_data)
view(electricity_survey)

f1 <- electricity_survey %>%
  select(respondent, TYPE, Value, B38) %>%
  filter(TYPE == 'general')

f1 <- f1 %>%
  group_by(B38) %>%
  summarise(totalWH = sum(Value))

ggplot(f1, aes(x = B38, y = totalWH)) + geom_col(fill = '#40B0A6') + labs(title = 'Total General
Electricity Consumption', subtitle = 'Based on Household Sizes in Victoria', y = 'Total Consumpt
ion (WH)', x = 'Household Size')
```



The barplot aims to portray the relationship between household size and total general electricity consumption from 1 April 2012 to 31 March 2014 in Victoria. Logically speaking, there is a positive correlation between household size and electricity consumption (greater household size = greater consumption). Interestingly, this is not the case here. As we can see, the total general electricity consumption of a household with only one person is approximately 25,000,000 WH, and the figure increases to approximately 75,000,000 WH as the household size went up to 2. However, this is followed by a massive drop as household size went to 3, and now the total consumption is slightly less than 25,000,000 WH. We continue seeing this pattern as total consumption reaches 100,000,000 WH at household size 4 (which is the household size with the greatest amount of total general electricity consumption), before again dropping to approximately 12,500,000 WH at household size 5. There are

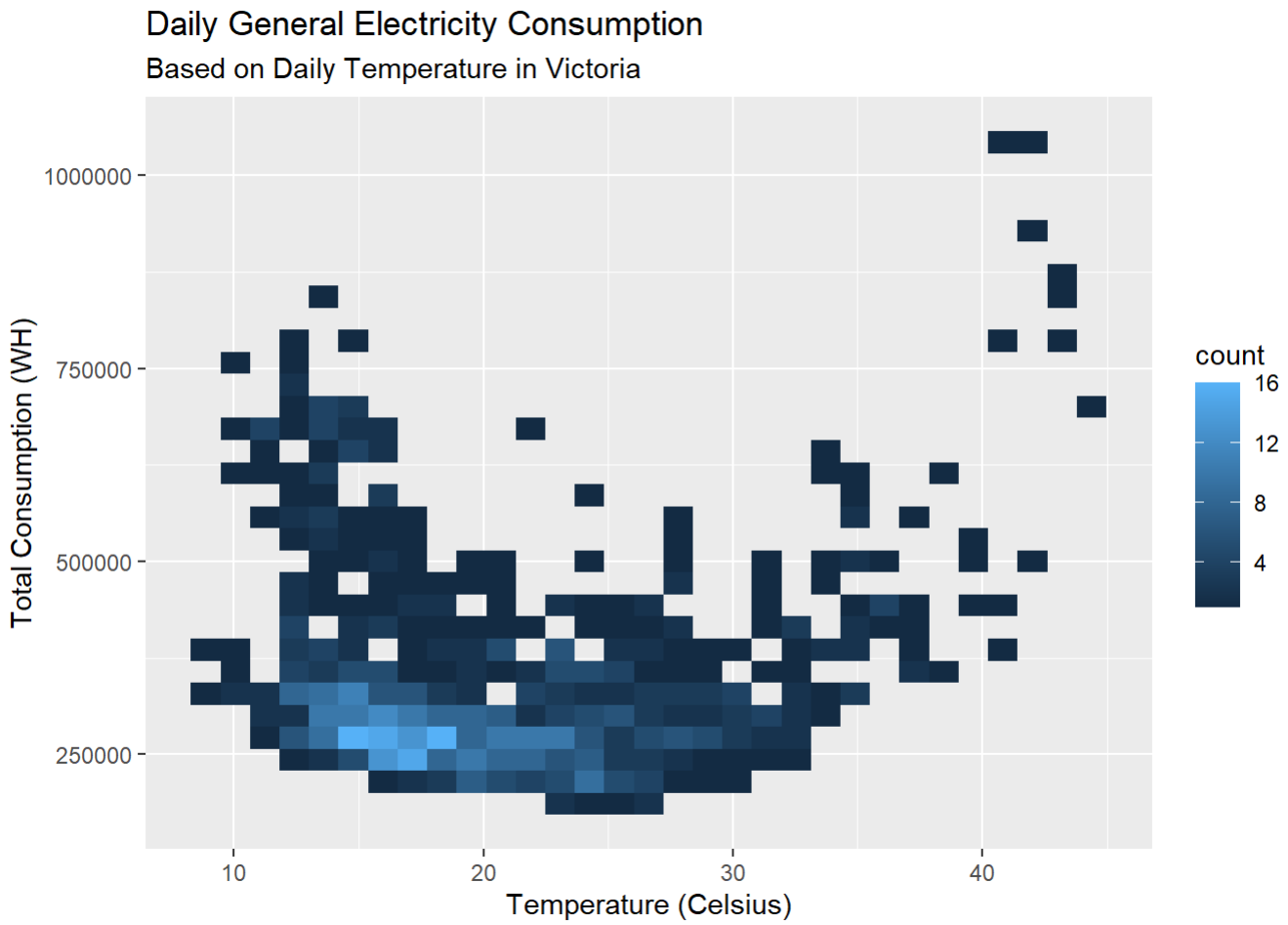
some possible reasons for this, such as majority of the households with sizes 3 and 5 living in rural or small regional centre, or the aging population of the households with sizes 3 and 5 (studies suggest that we use more electricity as we age).

Figure 2

```
e2 <- electricity_long %>%
  select(TYPE, Year, Month, Day, Value) %>%
  filter(TYPE == 'general')%>%
  group_by(Year, Month, Day)%>%
  summarize(totalWH = sum(Value))

electricity_weather <- full_join(e2, weather_data)
electricity_weather <- electricity_weather %>%
  select(Year, Month, Day, totalWH, `Maximum temperature (Degree C)`) %>%
  filter(!is.na(totalWH))
view(electricity_weather)

ggplot(electricity_weather, aes(x = `Maximum temperature (Degree C)`, y = totalWH)) + geom_bin2d
() + labs(title = 'Daily General Electricity Consumption', subtitle = 'Based on Daily Temperatur
e in Victoria', x = 'Temperature (Celsius)', y = 'Total Consumption (WH)')
```

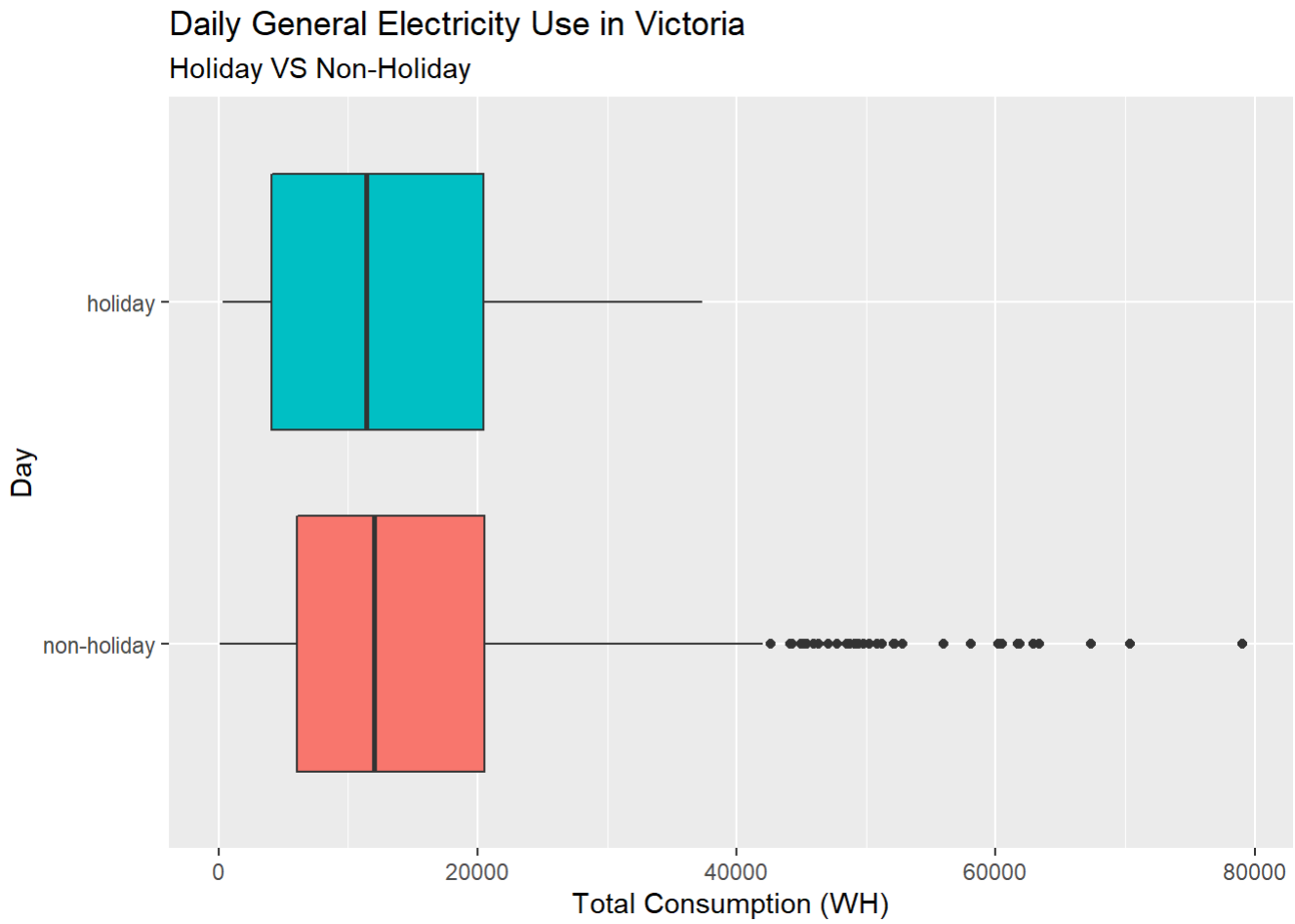


The heatmap visualizes the relationship between total daily general electricity consumption and daily temperature from 1 April 2012 to 31 March 2014 in Victoria. From a logical standpoint, cold temperature often leads to more electricity usage due to increased use of equipments such as heaters and electric boilers, while warm temperature

also translates to higher electricity consumption due to the increased use of air conditioners, fans, and so on. This is exactly what we can see from the plot. The plot is more “spread out” when daily temperature is below 15 Celsius, with majority of the points lying within the 500,000-750,000WH range, indicating high electricity consumption. The plot will then begin to flatten as it enters the 15-30 Celsius range, before again spreading out as daily temperature is above 30 Celsius. Also notice that when daily temperature is above 40 Celsius, electricity consumption is extremely high, with some points lying above 1,000,000WH. Although the relationship between general electricity consumption and daily temperature was fairly well explained, perhaps inclusion of some underlying factors, such as income or household size could help explain the plot, mainly the outliers, better.

Figure 3

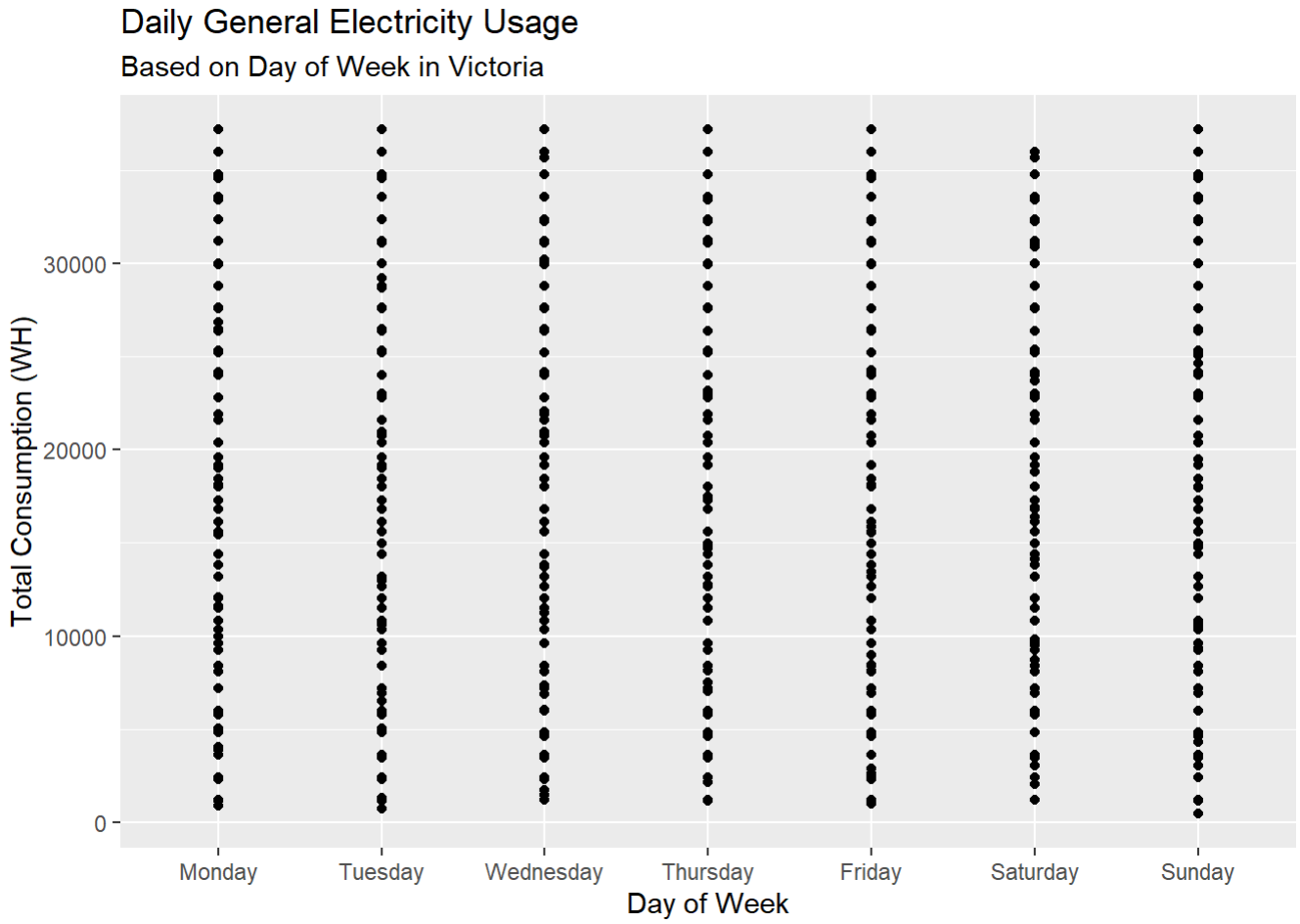
```
v1 <- important_dates$public_holidays
v1 <- factor(v1, levels = c(0,1), labels = c("non-holiday","holiday"))
important_dates1 <- important_dates %>%
  select(Year, Month, Day, public_holidays) %>% mutate(public_holidays = v1)
electricity_dates <- full_join(electricity_long,important_dates1)
electricity_dates <- electricity_dates %>%
  select(TYPE, Year, Month, Day, Value, public_holidays) %>%
  filter(TYPE == 'general')
n <- nrow(subset(electricity_dates, public_holidays == 'holiday'))
n <- n*2
electricity_dates <- head(electricity_dates,n)
electricity_dates <- electricity_dates %>%
  select(Year, Month, Day, Value, public_holidays) %>%
  group_by(Year, Month, Day, public_holidays)%>%
  summarize(totalWH = sum(Value))
ggplot(electricity_dates, aes(x=totalWH, y = public_holidays, fill = public_holidays)) + geom_bo
xplot(show.legend = FALSE) + labs(title = 'Daily General Electricity Use in Victoria', subtitle
 = 'Holiday VS Non-Holiday', x = 'Total Consumption (WH)', y = 'Day' ) + scale_color_viridis_c()
```



This boxplot aims to visualize the relationship between daily general electricity consumption and the 'type' of day (Holiday or Non-Holiday). To ensure fairness in analysis, an equal number of observations were taken for both 'holiday' and 'non-holiday'. Initially, it can be observed that the median of the 'non-holiday' boxplot is slightly ahead of 'holiday's, indicating that the average daily general electricity consumption is slightly greater during non-holidays. Next, while the 75th percentile (Q3) of both boxplots are arguably equal, the 25th percentile (Q1) of 'non-holiday' is ahead, suggesting that 75% of the non-holiday observations experience approximately 8,000 WH of daily electricity consumption, compared to holidays, where 75% of it experience 'only' approximately 4,000 WH worth of electricity consumption. Finally, while (again) the minimum daily electricity consumption is equal for both days, the maximum of 'non-holiday' is way ahead, lying slightly above 40,000 WH. In addition, there are some outliers in 'non-holiday', with one observation slightly below 80,000 WH. Thus, it can be concluded that non-holidays experience greater general electricity consumption. This is reasonable, considering people need to work, thus need more electricity, during these days.

Figure 4

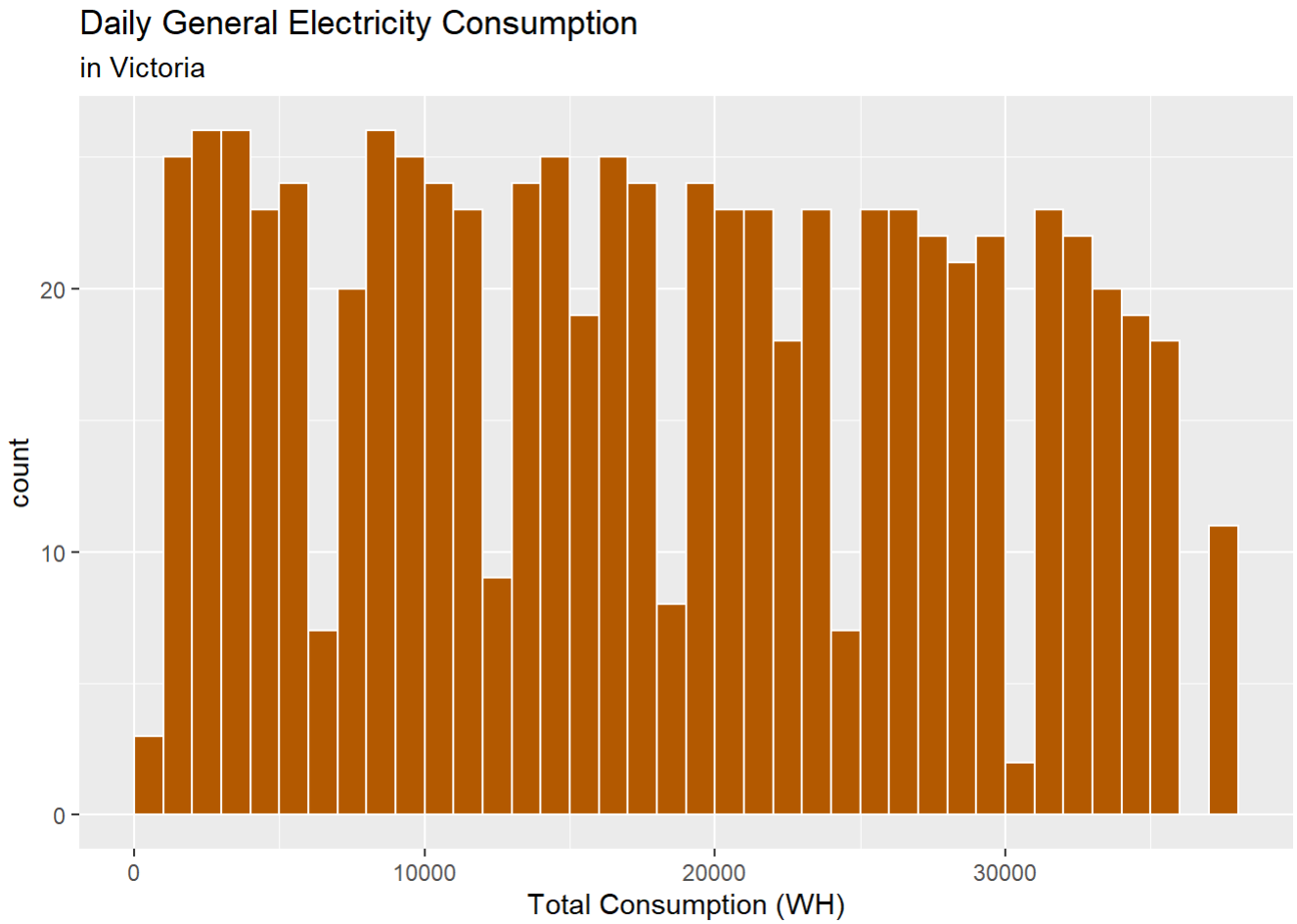
```
electricity_long <- electricity_long %>%
  select(TYPE, Year, Month, Day, Value) %>%
  filter(TYPE == 'general') %>%
  group_by(TYPE, Year, Month, Day) %>%
  summarize(totalWH = sum(Day))
electricity_dates1 <- full_join(electricity_long,important_dates)
electricity_dates1 <- electricity_dates1 %>%
  select(TYPE,Year, Month, day_of_week, totalWH) %>%
  filter((!is.na(TYPE)) & (!is.na(totalWH)))
ggplot(electricity_dates1, aes(y = totalWH, x = day_of_week)) + geom_point() + scale_x_discrete
(limits = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")) + lab
s(title = "Daily General Electricity Usage", subtitle = "Based on Day of Week in Victoria", y =
'Total Consumption (WH)', x = 'Day of Week')
```



This scatterplot visualizes the relationship between daily general electricity usage and the day of the week (i.e. Monday, Tuesday, etc.) from 1 April 2012 to 31 March 2014 in Victoria. First, notice that for most days, overplotting occurs within the 10000-20000WH range, although some overplotting slightly above 20000WH can also be seen (for instance on Wednesday). This suggests that the average electricity consumption on most days is within the 10000-20000WH range. Next, we can see that the tail/minimum value of Monday and Tuesday are similar, before increasing and remain constant on Wednesday, Thursday, Friday and Saturday, suggesting that on these days, people consume more electricity. The minimum/tail value then dropped again on Sunday. Finally, notice that Saturday has the smallest/shortest plot. One possible reason for this is that people spend more time outside (e.g. hanging out, going for a trip) than inside (at home) on Saturday, and therefore consume less electricity on Saturday than any other days.

Figure 5

```
ggplot(electricity_long, aes(x = totalWH)) + geom_histogram(fill = "#b35900" ,col = "white", bou
ndary = 0, binwidth = 1000) + labs(title = "Daily General Electricity Consumption", subtitle =
"in Victoria", x = "Total Consumption (WH)",y = "count")
```



The histogram aims to portray the number/count of daily general electricity usage from 1 April 2012 to 31 March 2014 in Victoria. The histogram shows a somewhat combed distribution, as we can see repeating peaks and troughs. This shape exists due to the fact that daily electricity consumption can take up any values and lie within any range, and therefore its distribution can take any shape. Next, a regular downward trend can be seen. This is reasonable, as it is illogical to assume that one person can consume high amount of electricity in a day. Therefore, there is a negative correlation between daily electricity consumption and count/frequency. In addition, notice the outlier on the 38000WH bin. A good explanation for this are days with extremely high/low temperature, where electricity usage rockets (as shown by the outliers on Figure 2). Finally, we can see that the 3000, 4000 and 9000WH bins have the highest frequency (around 27). This means the most frequent daily general electricity consumption observed take those values. Meanwhile, the least frequent count can be observed on the 31000WH bin (around 2).