BUS 316 – Project 2 – Business Analytics with R

Chaeyon Jang

2023-12-04

Load Packages

```
library(tidyverse)
library(fosdata)
library(ggplot2)
options(scipen=999)
```

Load Data

```
ev_cars <- ecars
```

Explore Data

The first step in answering the questions below is to explore the ecars dataset using the appropriate exploratory R functions. Use two R functions to explore the data. One of these functions is to show the structure of the dataset and provide a sampling of the data. For the second exploratory function you are to use a function that will output the first 6 rows of the ecars dataset.

We can use str(ev_cars) to get the structure and use either head(ev_cars, 6) or ev_cars %>% slice(1:6) to get the first 6 rows of the ecars dataset.

Based on your exploration write a brief description of the dataset. How many rows and how many columns does the ecars dataset have? Use R inline code to include the number of rows and columns in your description.

```
str(ev_cars)
```

```
## 'data.frame':    3395 obs. of  17 variables:
##  $ sessionId     : int  1366563 3075723 4228788 3173284 3266500 4099366 5084244 29484
36 3515913 8490014 ...
##  $ kwhTotal      : num  7.78 9.74 6.76 6.17 0.93 2.14 0.3 1.82 0.81 1.98 ...
##  $ dollars       : num  0 0 0.58 0 0 0 0 0 0 0 ...
##  $ created       : chr  "0014-11-18 15:40:26" "0014-11-19 17:40:26" "0014-11-21 12:0
5:46" "0014-12-03 19:16:12" ...
##  $ ended         : chr  "0014-11-18 17:11:04" "0014-11-19 19:51:04" "0014-11-21 16:4
6:04" "0014-12-03 21:02:18" ...
##  $ startTime     : int  15 17 12 19 20 14 15 20 17 18 ...
##  $ endTime       : int  17 19 16 21 21 15 15 21 18 18 ...
##  $ chargeTimeHrs : num  1.511 2.177 4.672 1.768 0.299 ...
##  $ weekday       : chr  "Tue" "Wed" "Fri" "Wed" ...
##  $ platform      : chr  "android" "android" "android" "android" ...
##  $ distance      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ userId        : int  35897499 35897499 35897499 35897499 35897499 35897499 3589749
9 35897499 35897499 35897499 ...
##  $ stationId     : int  582873 549414 129465 569889 414088 911231 920264 431796 13442
7 207262 ...
##  $ locationId    : int  461655 461655 461655 461655 566549 202527 461655 461655 62090
6 928191 ...
##  $ managerVehicle: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ facilityType  : Factor w/ 4 levels "Manufacturing",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ reportedZip   : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
head(ev_cars, 6)
```

```
##   sessionId kwhTotal dollars             created               ended startTime
## 1   1366563     7.78    0.00 0014-11-18 15:40:26 0014-11-18 17:11:04        15
## 2   3075723     9.74    0.00 0014-11-19 17:40:26 0014-11-19 19:51:04        17
## 3   4228788     6.76    0.58 0014-11-21 12:05:46 0014-11-21 16:46:04        12
## 4   3173284     6.17    0.00 0014-12-03 19:16:12 0014-12-03 21:02:18        19
## 5   3266500     0.93    0.00 0014-12-11 20:56:11 0014-12-11 21:14:06        20
## 6   4099366     2.14    0.00 0014-12-12 14:38:44 0014-12-12 15:04:04        14
##   endTime chargeTimeHrs weekday platform distance   userId stationId locationId
## 1      17     1.5105556     Tue  android       NA 35897499    582873     461655
## 2      19     2.1772222     Wed  android       NA 35897499    549414     461655
## 3      16     4.6716667     Fri  android       NA 35897499    129465     461655
## 4      21     1.7683333     Wed  android       NA 35897499    569889     461655
## 5      21     0.2986111     Thu  android       NA 35897499    414088     566549
## 6      15     0.4222222     Fri  android       NA 35897499    911231     202527
##   managerVehicle facilityType reportedZip
## 1              0        R & D           0
## 2              0        R & D           0
## 3              0        R & D           0
## 4              0        R & D           0
## 5              0        R & D           0
## 6              0        R & D           0
```

```
ev_cars %>%
  slice(1:6)
```

```
##   sessionId kwhTotal dollars             created               ended startTime
## 1   1366563     7.78    0.00 0014-11-18 15:40:26 0014-11-18 17:11:04        15
## 2   3075723     9.74    0.00 0014-11-19 17:40:26 0014-11-19 19:51:04        17
## 3   4228788     6.76    0.58 0014-11-21 12:05:46 0014-11-21 16:46:04        12
## 4   3173284     6.17    0.00 0014-12-03 19:16:12 0014-12-03 21:02:18        19
## 5   3266500     0.93    0.00 0014-12-11 20:56:11 0014-12-11 21:14:06        20
## 6   4099366     2.14    0.00 0014-12-12 14:38:44 0014-12-12 15:04:04        14
##   endTime chargeTimeHrs weekday platform distance   userId stationId locationId
## 1      17     1.5105556     Tue  android       NA 35897499    582873     461655
## 2      19     2.1772222     Wed  android       NA 35897499    549414     461655
## 3      16     4.6716667     Fri  android       NA 35897499    129465     461655
## 4      21     1.7683333     Wed  android       NA 35897499    569889     461655
## 5      21     0.2986111     Thu  android       NA 35897499    414088     566549
## 6      15     0.4222222     Fri  android       NA 35897499    911231     202527
##   managerVehicle facilityType reportedZip
## 1              0        R & D           0
## 2              0        R & D           0
## 3              0        R & D           0
## 4              0        R & D           0
## 5              0        R & D           0
## 6              0        R & D           0
```

```
nrow(ev_cars)
```

```
## [1] 3395
```

```
ncol(ev_cars)
```

```
## [1] 17
```

The ev_cars dataset contains data that describes information on EV charging sessions; it measures charging time, facility type, charging by the week, price of each charging session, and more. The ev_cars dataset has 17 columns and 3395 rows.

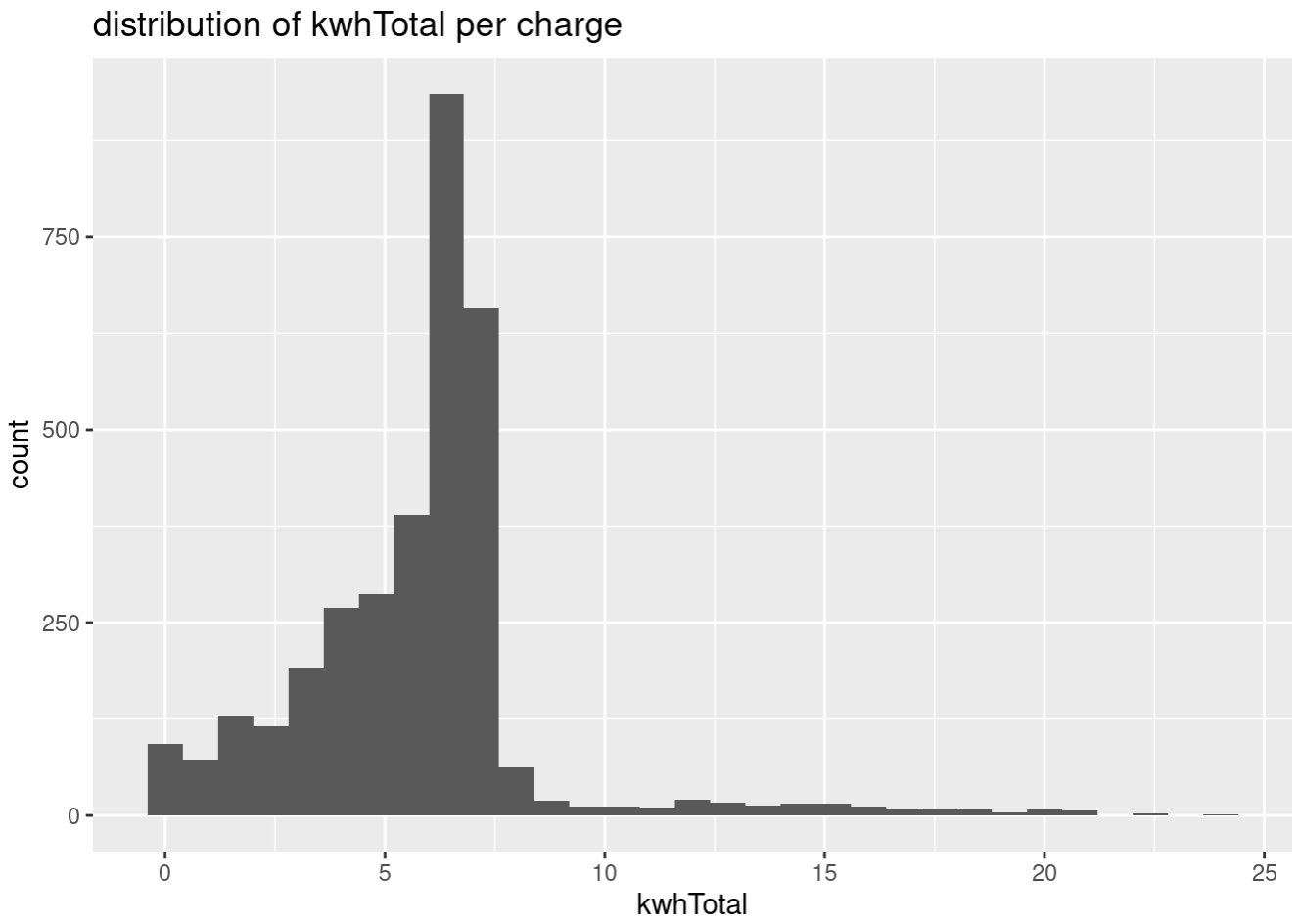
Question 1

Explore the distribution of data for the continuous variables of interest:

kwhTotal, chargeTimeHrs, distance, dollars

Use histograms to discuss the shape of each distribution.

```
ggplot(data = ev_cars, mapping = aes(x = kwhTotal)) +
  geom_histogram(binwidth = 0.80) +
  labs(title = "distribution of kwhTotal per charge")
```

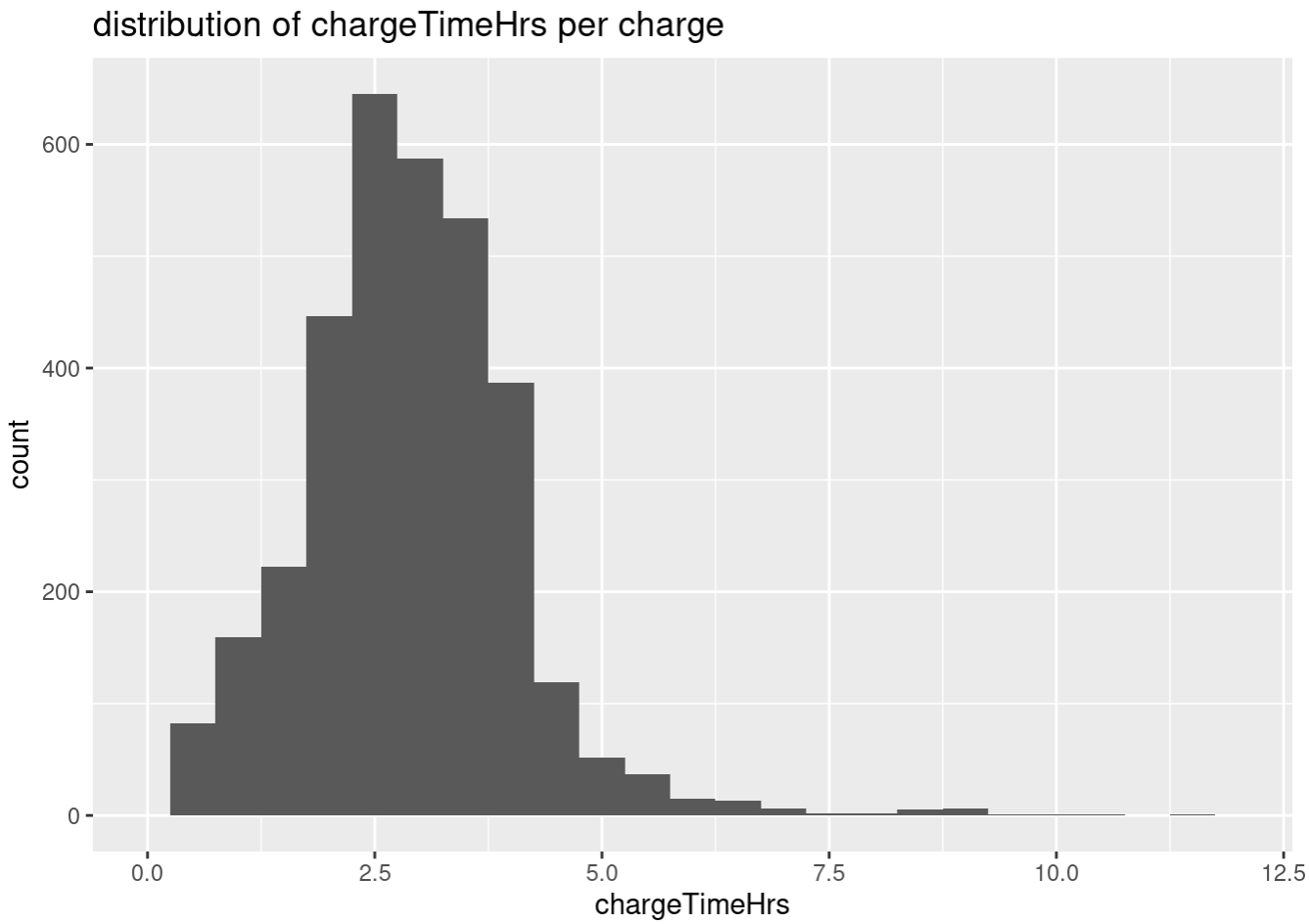


The distribution of the kwhTotal is right-skewed, with a unimodal distribution. It seems that the kwhTotal with the highest concentration is around 6-7 kwhs (kilowatthours). The count of kwhTotal dramatically decreases after the 6-7 kwh peak, signifying the inefficiency and dangers of charging the EV to a high kwh.

```
ggplot(data = ev_cars, mapping = aes(x = chargeTimeHrs)) +
  geom_histogram(binwidth = 0.5) +
  xlim(0, 12) +
  labs(title = "distribution of chargeTimeHrs per charge")
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```

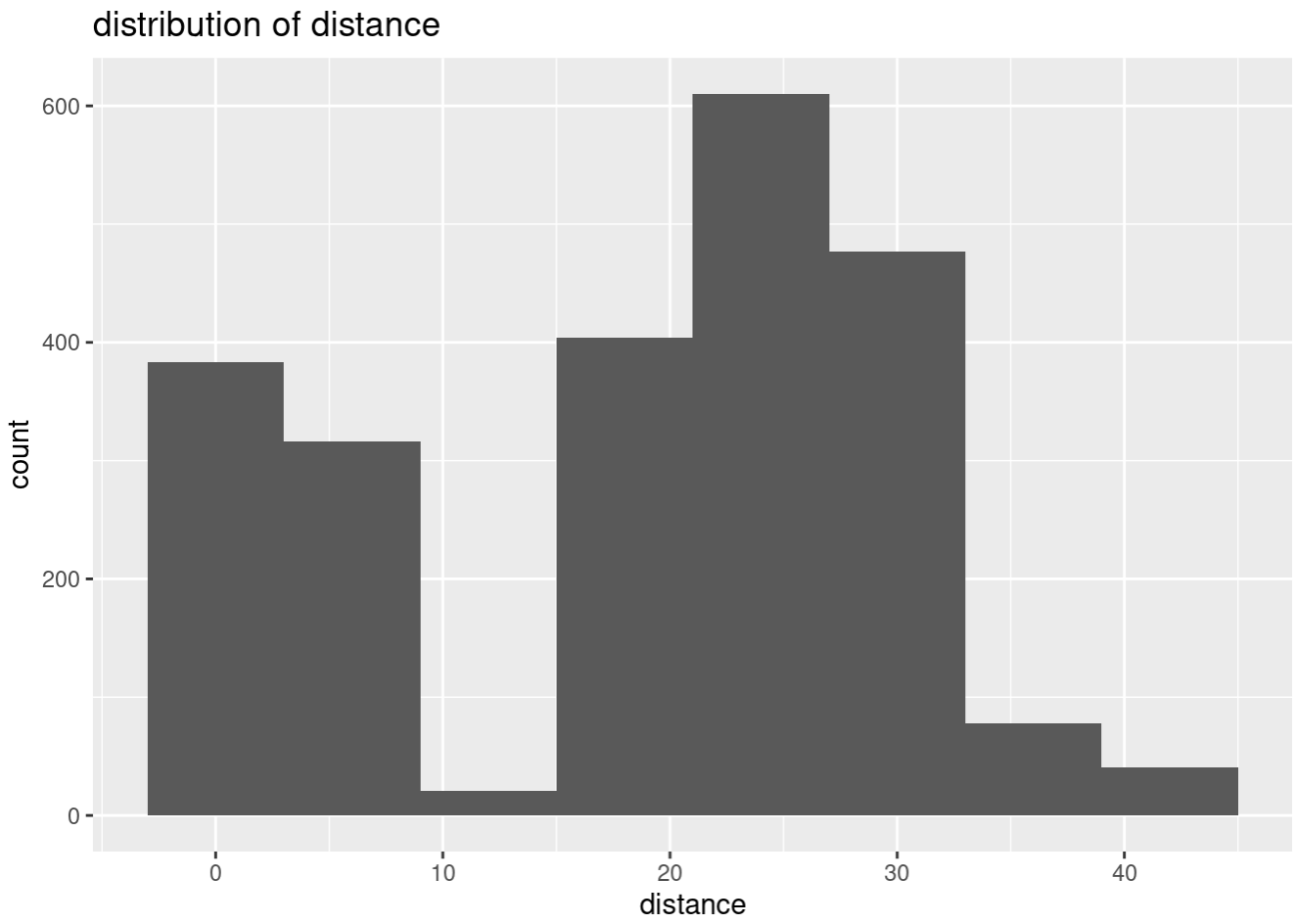
```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



The chargeTimeHrs is similar to the kwhTotal data, where the histogram is unimodal and right-skewed. However, in comparison to the kwhTotal distribution, the chargeTimeHrs have a more symmetrical shape, where there is no extreme decrease or increase in the pattern of data provided. Looking at the histogram, most EV owners charge their vehicles for aronud 2.5-3 hours.

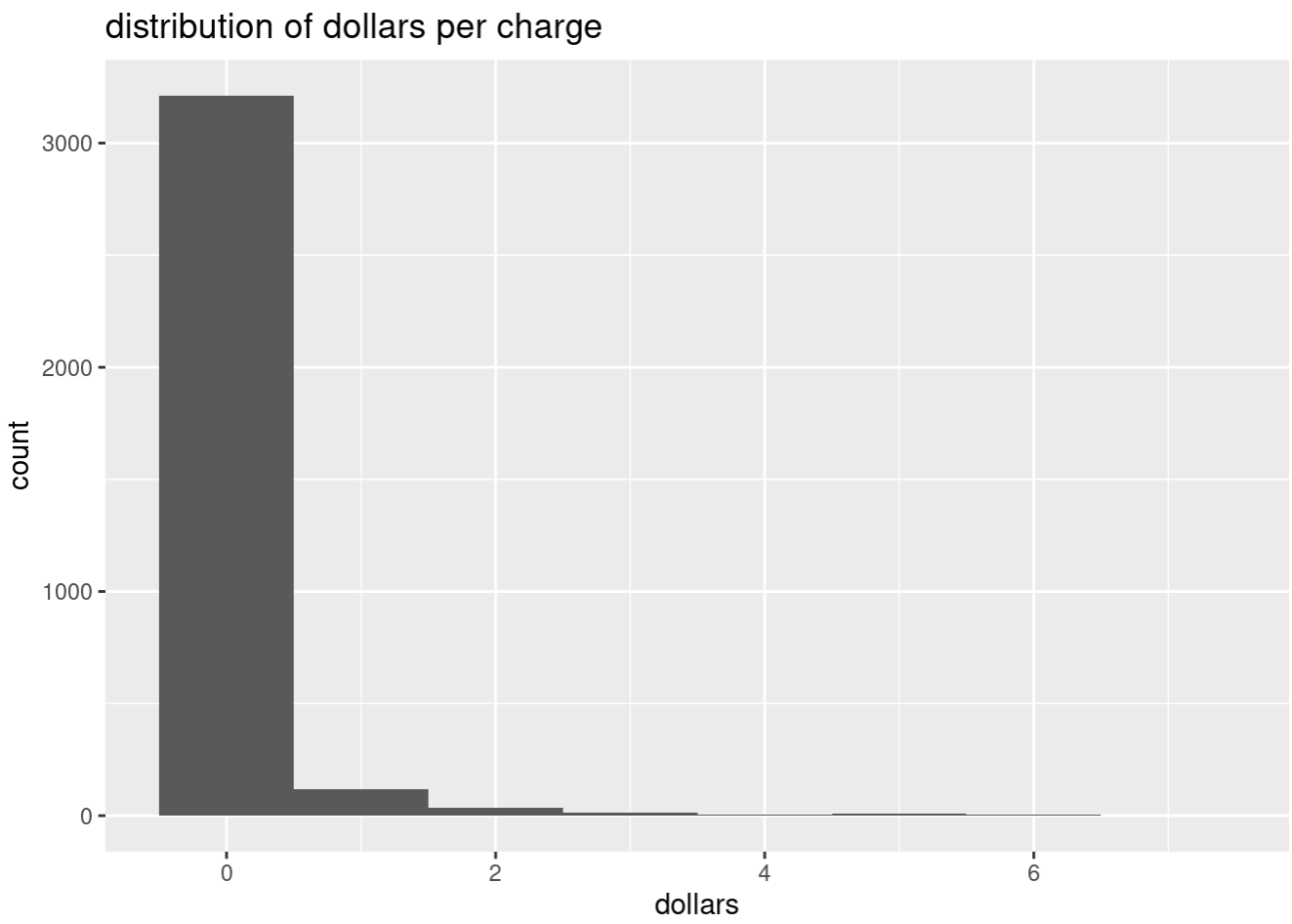
```
ggplot(data = ev_cars, mapping = aes(x = distance)) +
  geom_histogram(binwidth = 6) +
  labs(title = "distribution of distance")
```

```
## Warning: Removed 1065 rows containing non-finite values (`stat_bin()`).
```



The histogram showing the distribution of distance is very spread out. There is no symmetry to the graph, nor is there any noticeable pattern. One thing, however, is that there seems to be concentrations in some distances, such as 0-7, 20, 23, and ~34. These spikes could describe that employees typically live around within 8 miles, around 20/23/34 miles away from their workplace, where the chargers are.

```
ggplot(data = ev_cars, mapping = aes(x = dollars)) +
  geom_histogram(binwidth = 1) +
  labs(title = "distribution of dollars per charge")
```



Looking at the distribution of dollars, it seems that charging sessions are mainly free, with the exception of a few stations that charge $1-$5. The histogram is heavily skewed right, showing that most employees get free charging services for their EVs.

Question 2

Next we will examine the distribution of the categorical variables of interest:

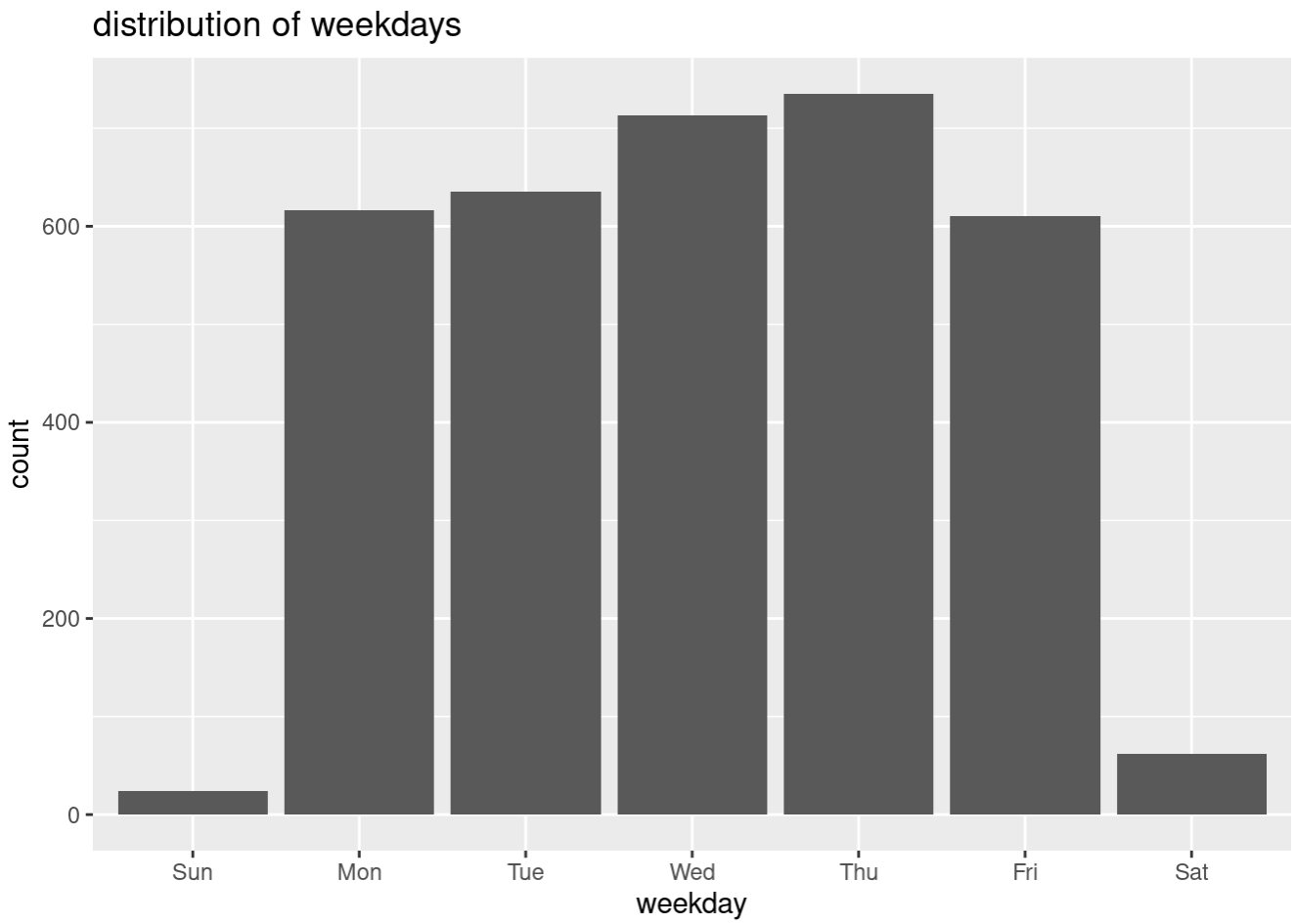
weekday, platform, facilityType

Use bar plots for this analysis, and describe each distribution.

To make the bar plot order the days of the week, we modify the tidyverse to recreate the weekdays in order.

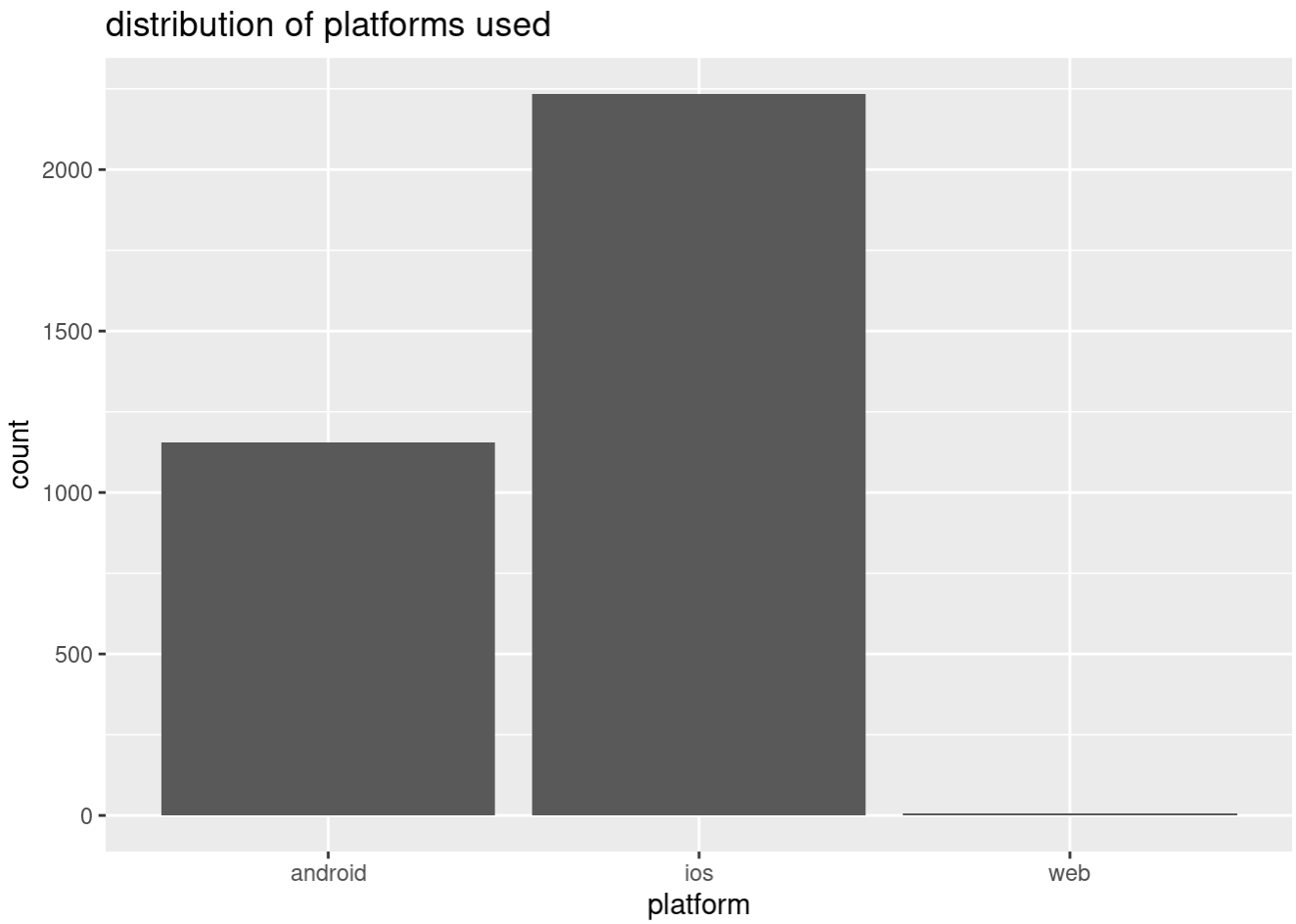
```
days <- c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat")
ev_cars <- ev_cars %>%
  mutate(weekday = factor(weekday, levels = days))
```

```
ggplot(data = ev_cars, mapping = aes(x = weekday)) +
  geom_bar() +
  labs(title = "distribution of weekdays")
```



Looking at the bar graph, it seems that most employees charge their vehicles at work during the weekdays, which makes sense because most workers are off during the weekend.

```
ggplot(data = ev_cars, mapping = aes(x = platform)) +
  geom_bar() +
  labs(title = "distribution of platforms used")
```



This graph shows the distribution of charging platforms. The ios platform is used by most employees, followed by the android. This data shows the possible preference of the ios platform, perhaps denotating that the ios platform charges the best or is a better cost-efficient charging platform.

```
ggplot(data = ev_cars, mapping = aes(x = facilityType)) +
  geom_bar() +
  labs(title = "distribution of facilityType")
```

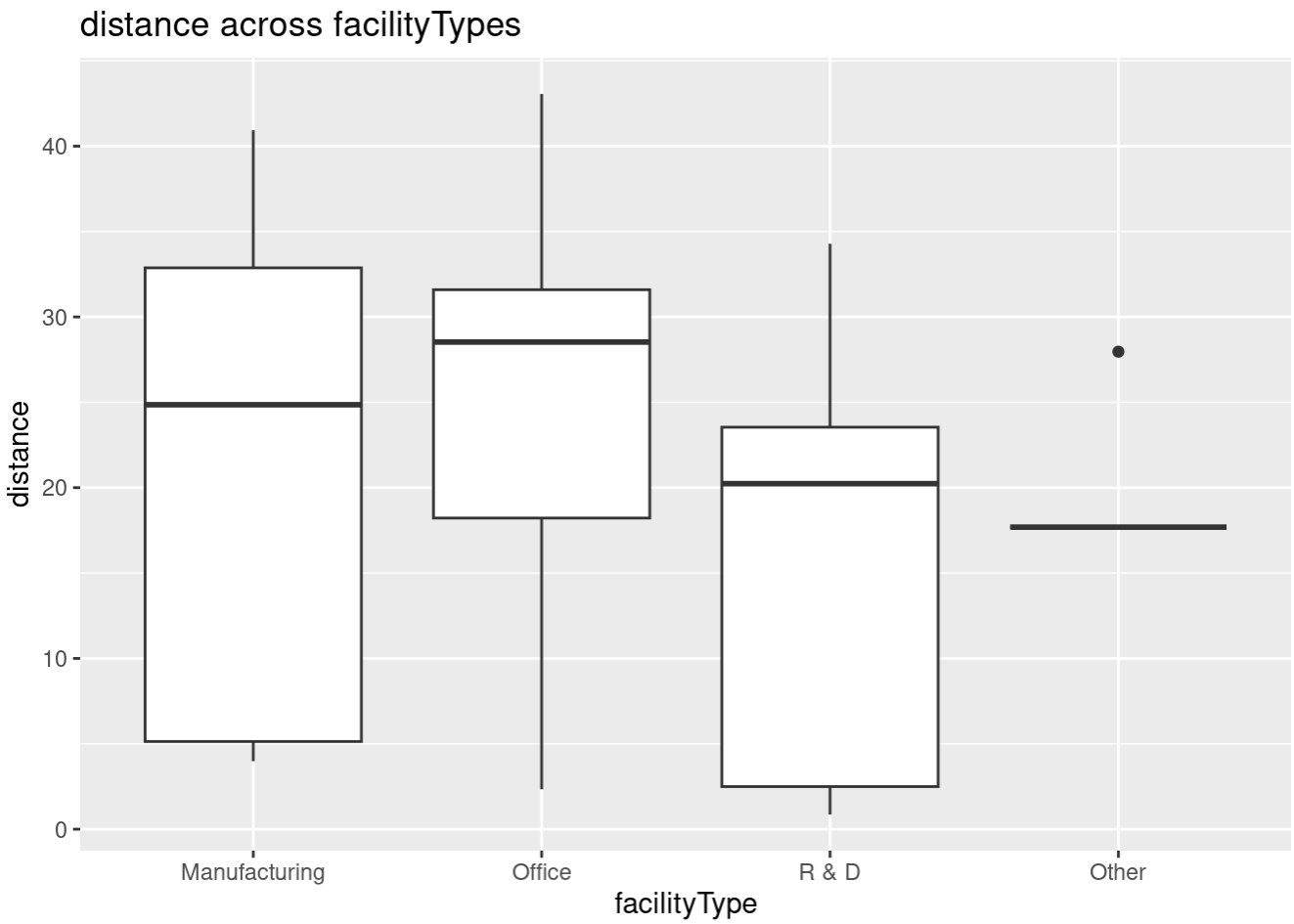
The bar graph shows the distribution of facilityType, where the charging stations are placed. The overall shape of the graph is unimodal, with most charging stations being placed at the R&D facility.

Question 3

Boxplots are commonly used visualizations to get a feel for the median and spread of a variable, especially when you want to compare variations across groups in contained categorial variables. Create a boxplot to compare distance across facilityType. Discuss your findings.

```
ggplot(data = ev_cars, mapping = aes(x = facilityType, y = distance)) +
  geom_boxplot() +
  labs(title = "distance across facilityTypes")
```

```
## Warning: Removed 1065 rows containing non-finite values (`stat_boxplot()`).
```

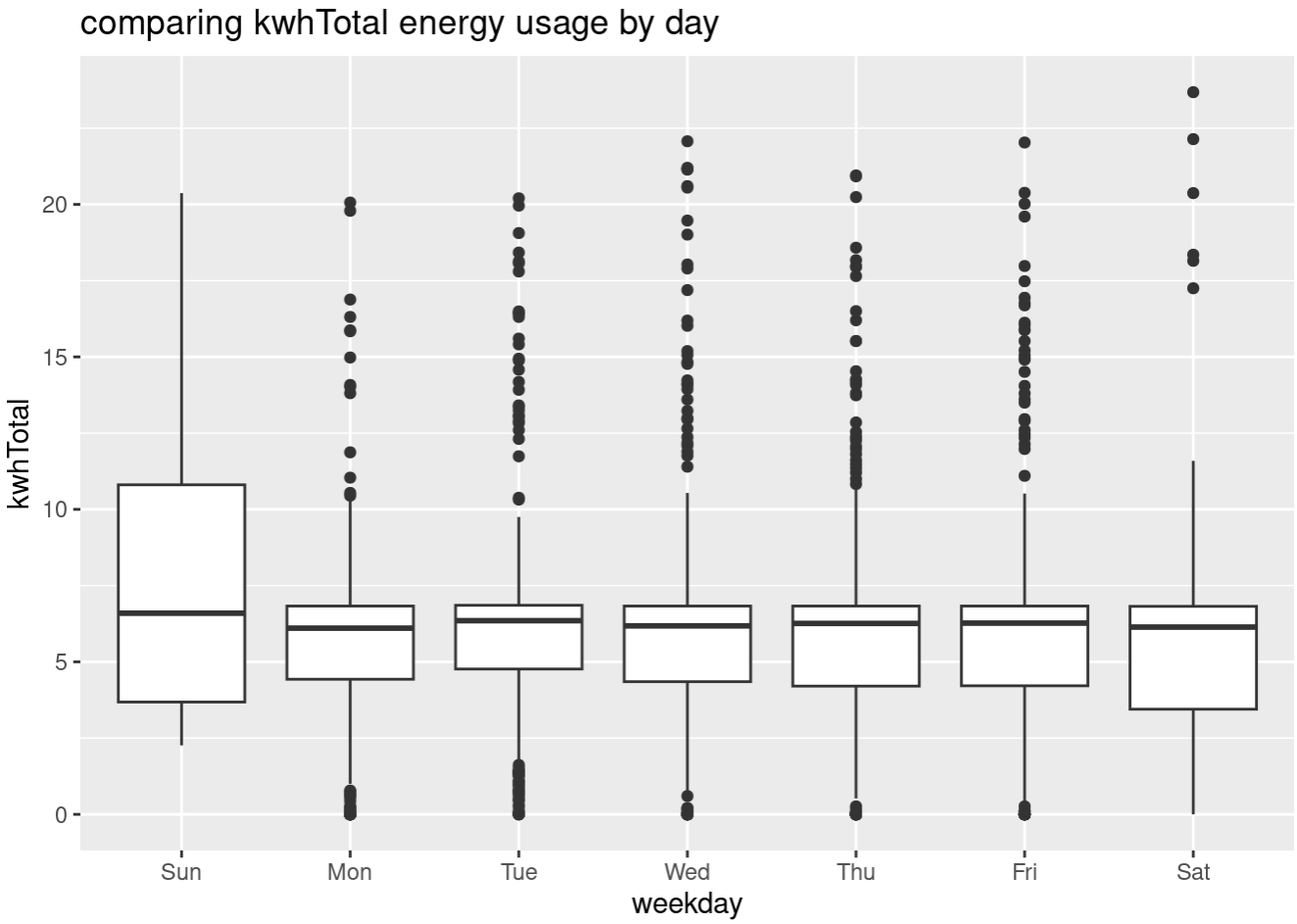


Looking at the boxplots comparing facilityType and distance, it seems that there is much more spread in distance driven to the manufacturing facility. Looking at the office facility, the majority of employees working in the office seem to have a smaller interquartile range, where majority of the employees commuting to the office has the highest median distance but a more consolidated IQR, meaning that on average, majority of the workers live around 20-35 miles away. However, the data for commuting to the office seems to have the largest total range, with some employees living less than 5 miles away to some living over 45 miles away. R&D seems to have a lot of spread in its IQR as well, but a lower median distance driven to the workplace.

Question 4

Next compare kwhTotal energy usage by day. Use boxplot for this analysis. How does energy usage vary by day?

```
ggplot(data = ev_cars, mapping = aes(x = weekday, y = kwhTotal)) +
  geom_boxplot() +
  labs(title = "comparing kwhTotal energy usage by day")
```



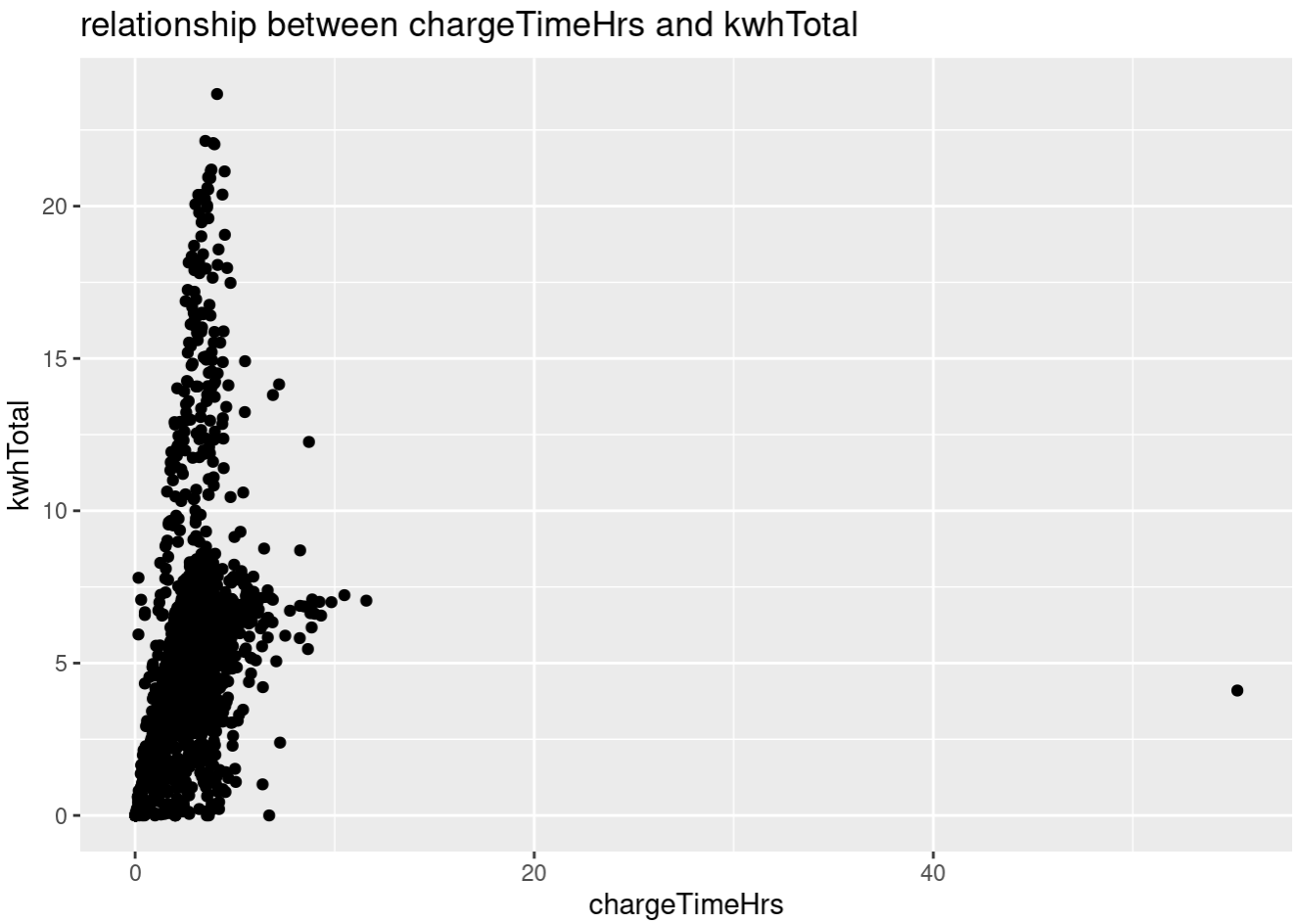
It seems that majority of the data points measuring the kwhTotal used throughout the workweek is consolidated to around 4.5-7 kwhTotal (IQR) and the median being around 6-7 kwhTotal, with a significant amount of outliers. There is a lot more spread in the data when looking at the weekends, which could be explained by less data points (from employees working less on the weekends) making the data seem more susceptible to changes in kwhTotal.

Question 5

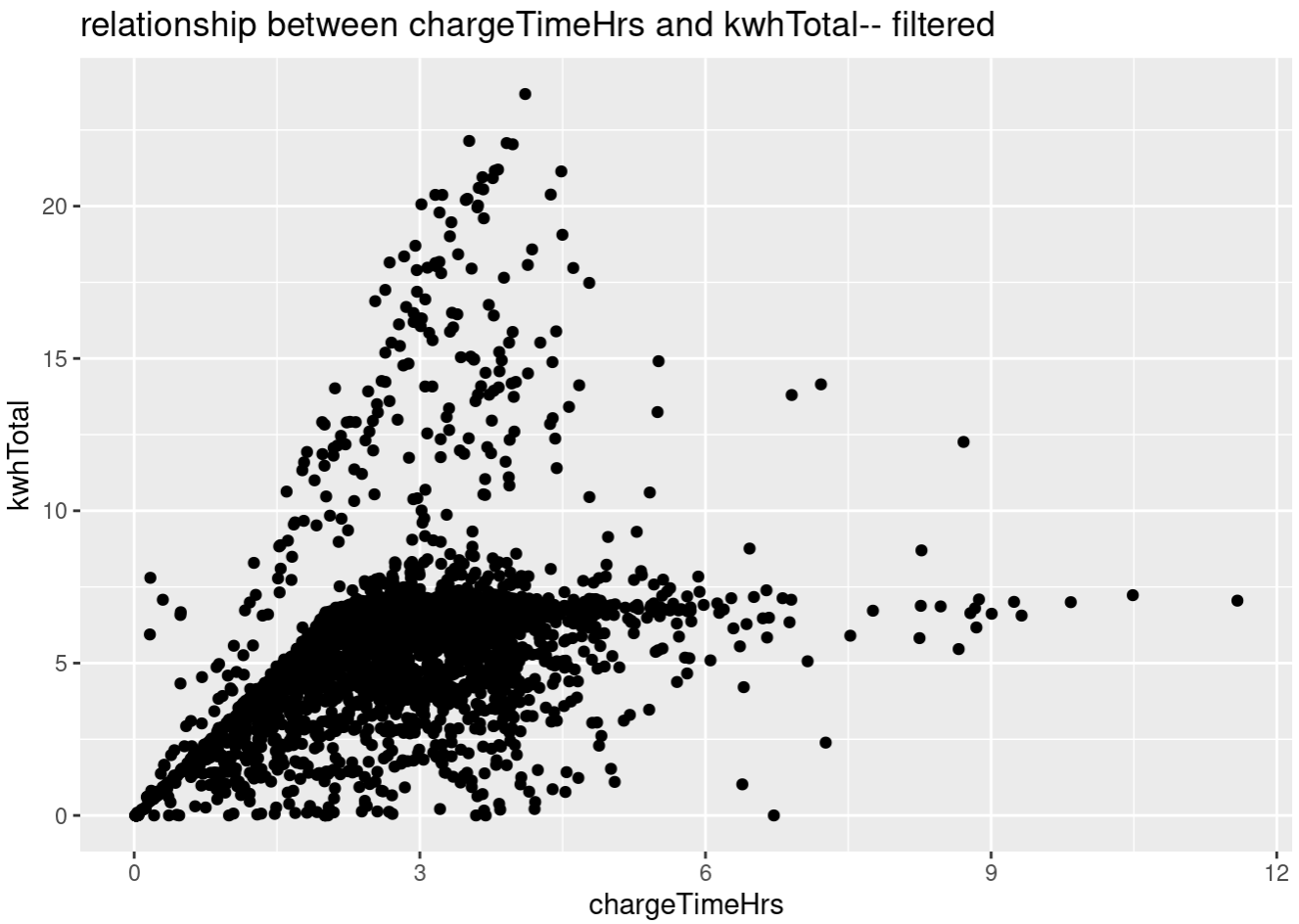
Explore the relationship between time at the charging station (chargeTimeHrs) and power usage (kwhTotal) for electric vehicles.

A first attempt at a scatterplot reveals that there is one data point with a car that charged for 55 hours. All other charge times are much shorter, therefore filter out this one outlier by including only change times less than 24 hours and regenerate the plot.

```
ggplot(data = ev_cars, mapping = aes(x = chargeTimeHrs, y = kwhTotal)) +
  geom_point() +
  labs(title = "relationship between chargeTimeHrs and kwhTotal")
```




```
b <- ev_cars %>%
  filter(chargeTimeHrs < 24)
ggplot(b, mapping = aes(x = chargeTimeHrs, y = kwhTotal)) +
  geom_point() +
  labs(title = "relationship between chargeTimeHrs and kwhTotal-- filtered")
```



Describe the relationship. Does anything seem unusual? What does the step line of dots on the left side of the represent?

Looking at the scatter plots, there is a slight positive correlation where as chargeTimeHours increases, kwhTotal increases. However, there is a strong concentration of points where consumers tend to charge an average of 0-4 hours and get around 5-7 kwhTotal during the 0-4 hour charge time.

Something that looks unusual is the wide spread of data in terms of the amount of kwhTotal a car gets per the same hours of charge; for 3 hours of charge, on average, each car gets around 3-8 kwhTotal, but some cars get up to 20 kwhTotal.

The step line of dots on the left side of the graph represents the concentration of consumers who charge for 0-4 hours and n average, get around 0-7 kwhTotal. The slow flattening of the step shows that it becomes less efficient to charge for more than a certain amount of hours because the amount of kwhTotal the EV gets is not as beneficial relative to the time that must be waited.

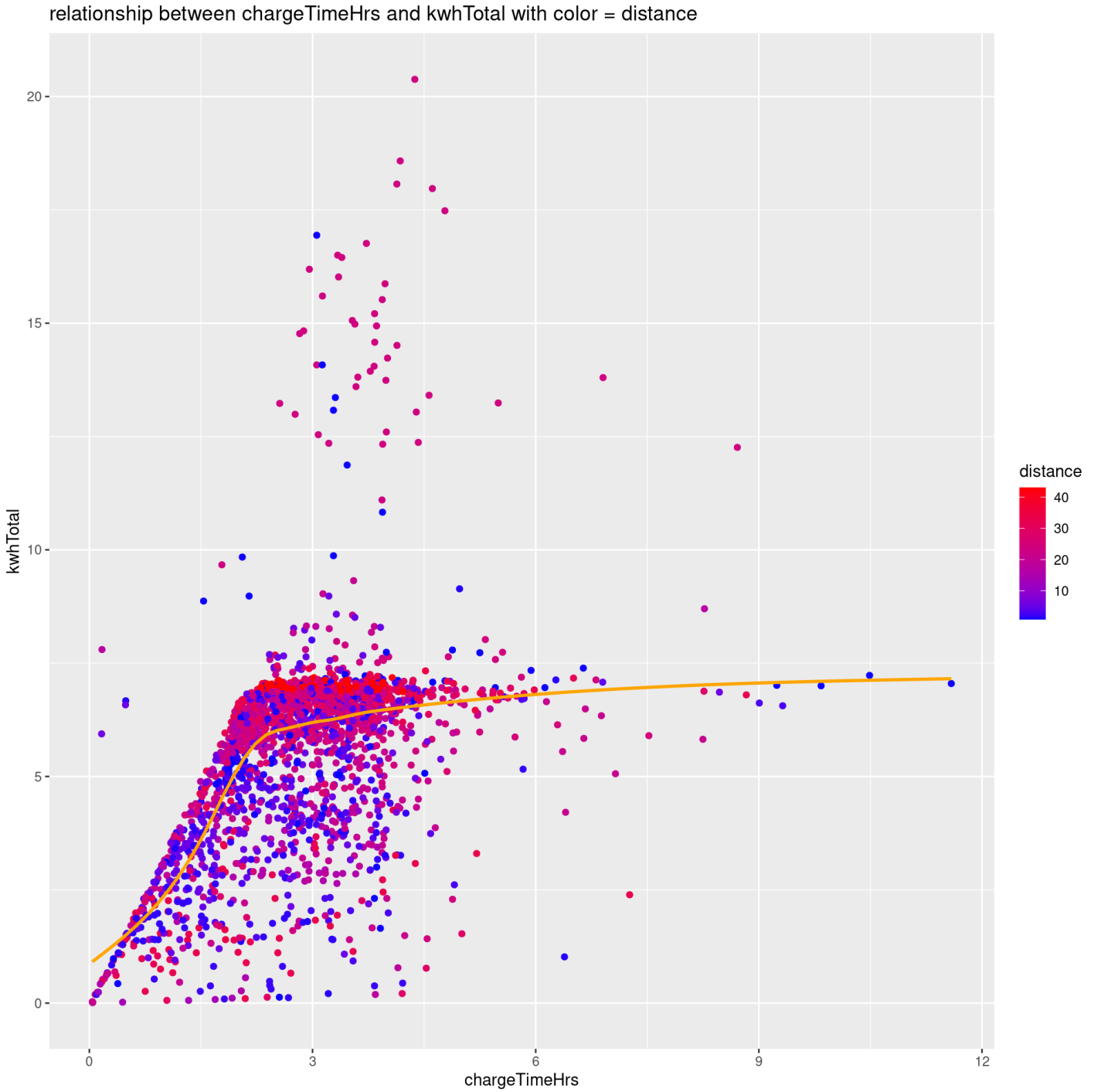
Question 6

Let's explore the distance variable as it relates to charging time and energy usage. Charging stations in this data set are generally workplaces, and the distance variable reports the distance to the drivers home. Only about 1/3 of the records in ecars contain valid distance data, so you need to filter for those. Basically, filter out the NAs. Next, assign distance to the color aesthetic and regenerate the scatterplot.

```
b <- ev_cars %>%
  filter(chargeTimeHrs < 24, !is.na(distance))
```

```
ggplot(b, mapping = aes(x = chargeTimeHrs, y = kwhTotal, color = distance)) +
  scale_color_continuous(low = "blue", high = "red") +
  geom_point() +
  geom_smooth(color = "orange", se = FALSE) +
  labs(title = "relationship between chargeTimeHrs and kwhTotal with color = distance")
```

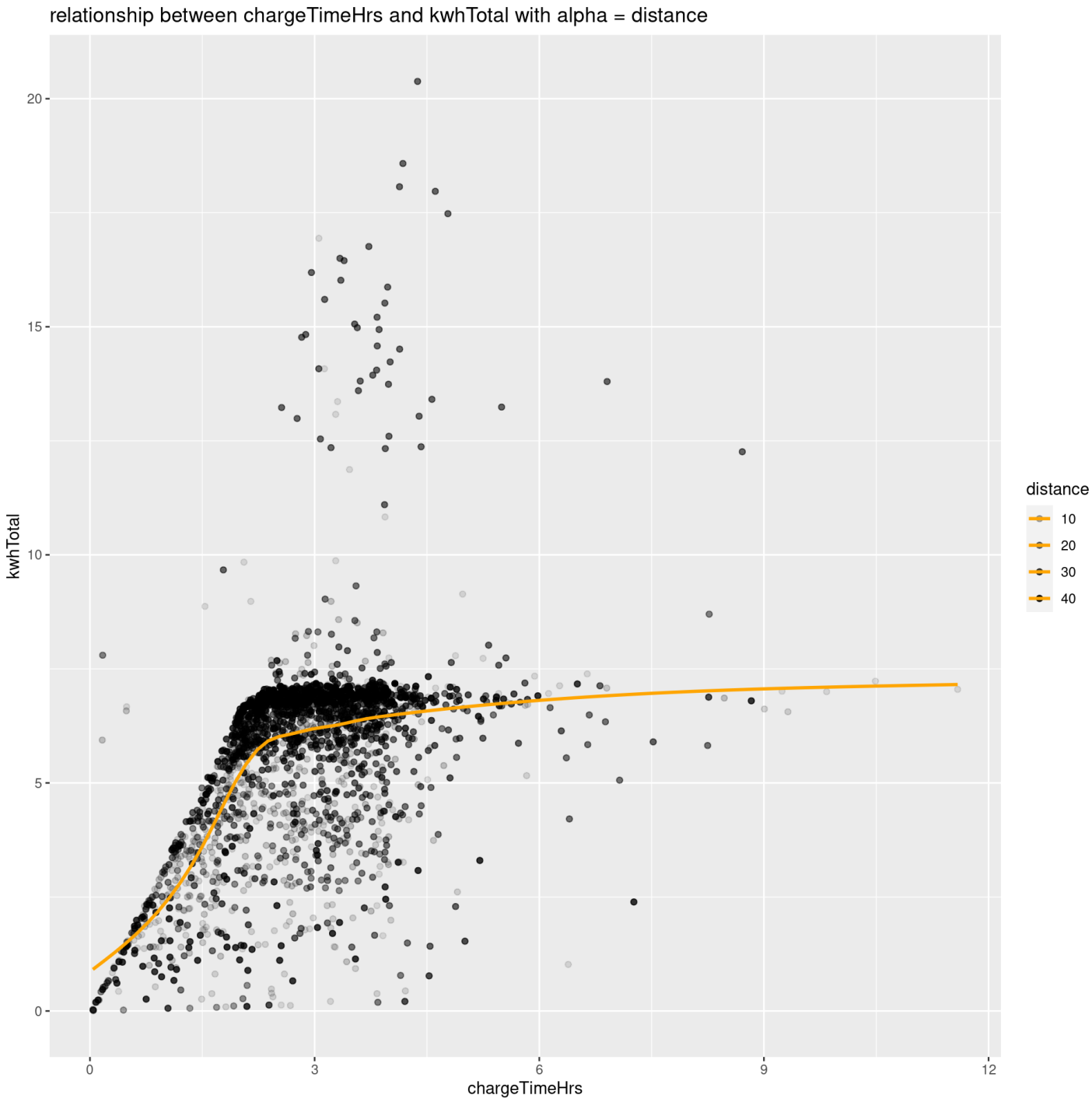
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
ggplot(b, mapping = aes(x = chargeTimeHrs, y = kwhTotal, alpha = distance)) +
  scale_color_continuous(low = "blue", high = "red") +
  geom_point() +
  geom_smooth(color = "orange", se = FALSE) +
  labs(title = "relationship between chargeTimeHrs and kwhTotal with alpha = distance")
```

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

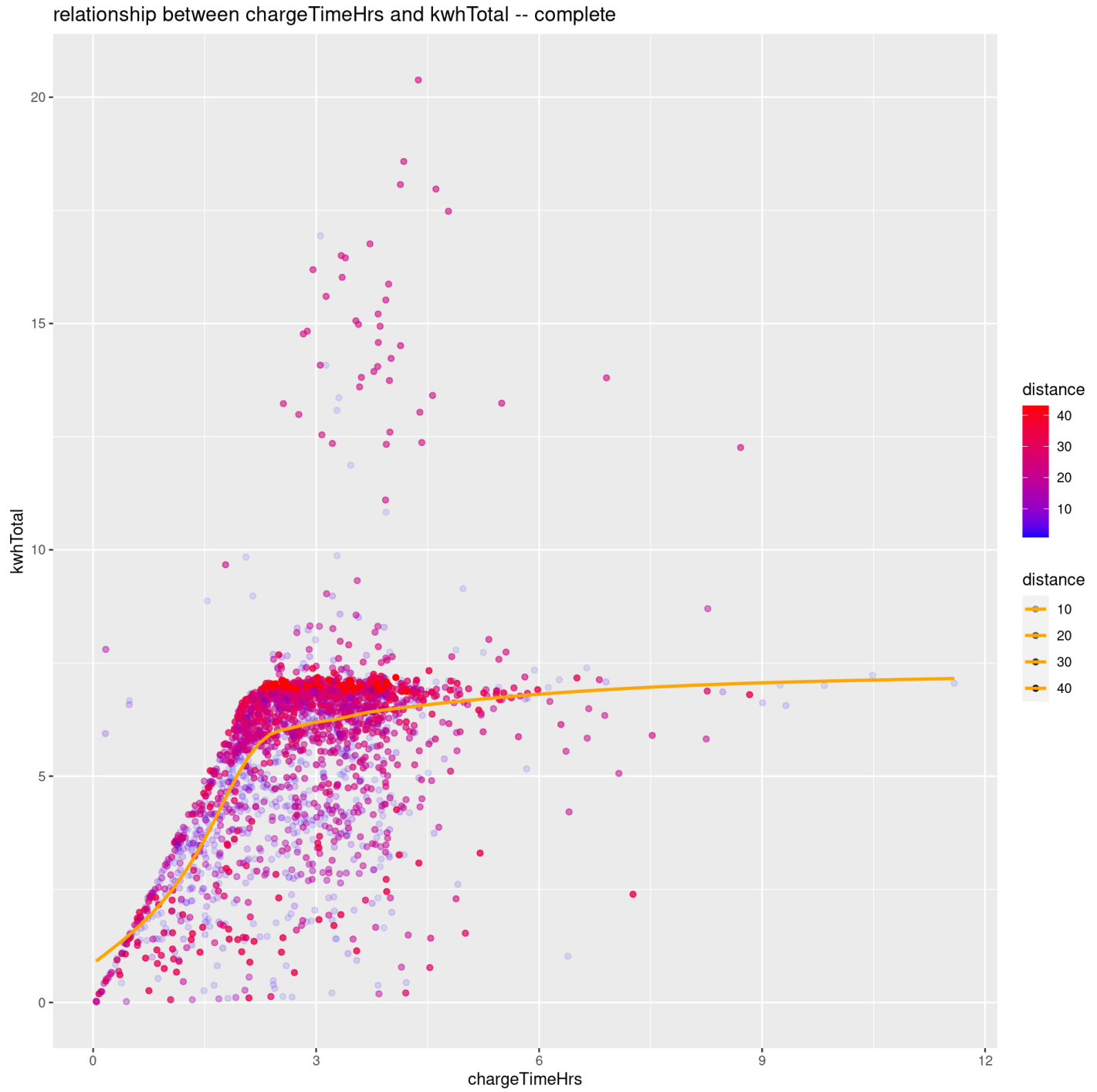
```
## Warning: The following aesthetics were dropped during statistical transformation: alp
ha
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
ggplot(b, mapping = aes(x = chargeTimeHrs, y = kwhTotal, color = distance, alpha = dista
nce)) +
  scale_color_continuous(low = "blue", high = "red") +
  geom_point() +
  geom_smooth(color = "orange", se = FALSE) +
  labs(title = "relationship between chargeTimeHrs and kwhTotal -- complete")
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: alp
ha
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



The results look cluttered, therefore also assign distance to the alpha aesthetic, which will make shorter distances partially transparent. To help interpret the results add a trend line to the chart and change the color of the line to orange. Regenerate the scatterplot. Also configure the chunk options to make this a larger chart when it is knitted.

How does the plot look without filtering away the invalid data? How does it look without alpha transparency? Any interesting findings in this scatterplot that relate to distance?

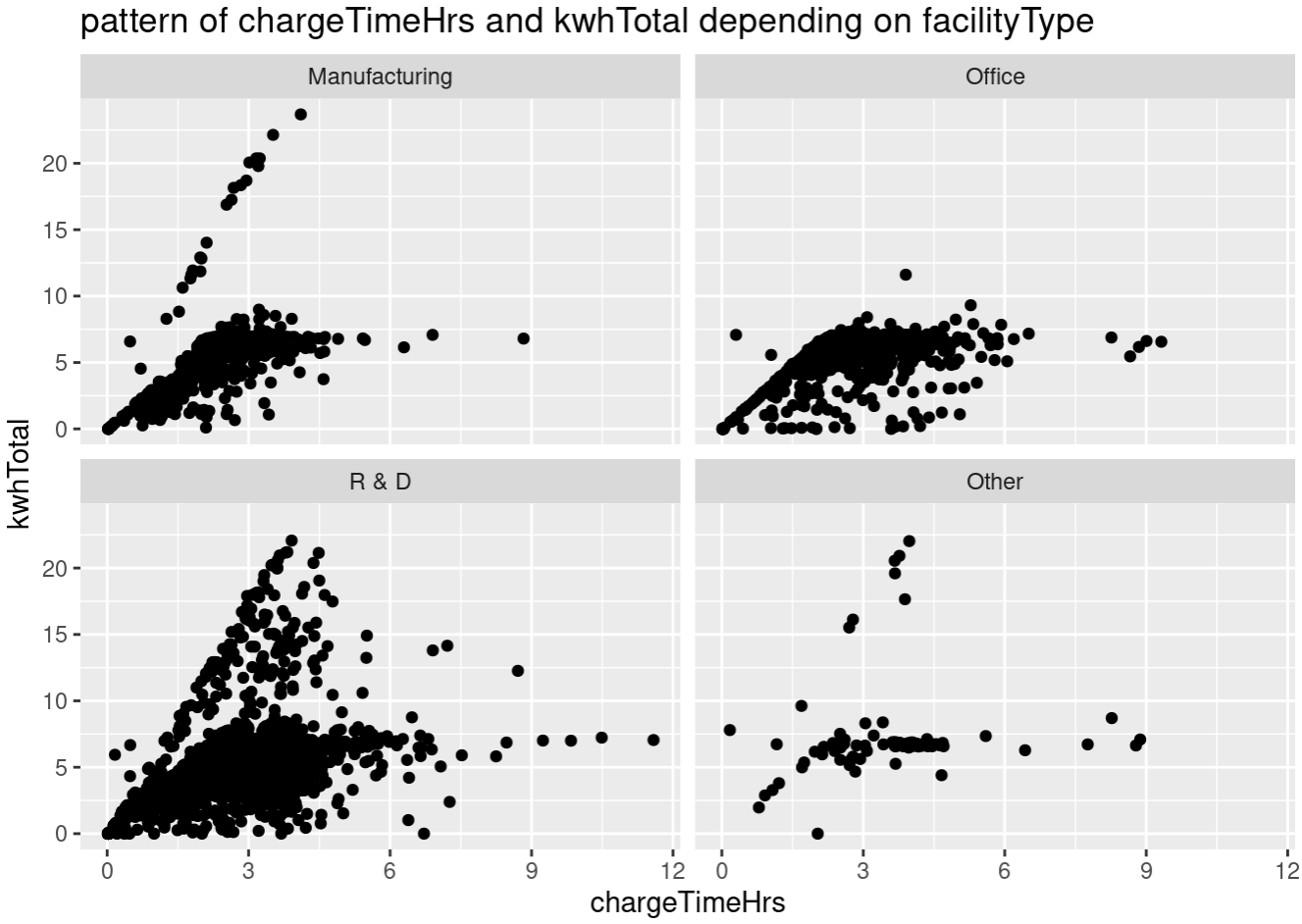
Looking at the first scatterplot in question 5, where there is no filter, the data is very cluttered and hard to see. The only information we can gather from that scatterplot is the general trend of chargeTimeHrs and kwhTotal. Without the alpha transparency, it is difficult to see the relationship between the chargeTimeHrs and kwhTotal relative to distance. The coloring of the first plot in question 6 shows the spread of distance data, but is hard to differentiate distance because of the overlapping of points. Therefore, by adding the alpha, the graph can show the shorter distances with a more transparent data point, with longer distances being darker.

An interesting pattern to note is that the employees that drive the farthest distance (non-transparent red) fall just above the trend line or are significantly below the line in kwhTotal, but the data points that are way above the trend line in regards to kwhTotal tend to be driving shorter distances. The chargeTimeHrs seem to level out regardless of distance driven, but the kwhTotal seems to vary across the scatterplot, which could be credited to the employee preference for charging or the platform/charging station efficiency.

Question 7

Does the usage pattern for electric car charging depend on the type of facility where the charging station is located? Use a facet wrap and the plot from Question 5 where you removed the one outlier to plot total KWH as a function of charging time in four scatterplots, one for each of the four facility types.

```
b <- ev_cars %>%
  filter(chargeTimeHrs < 24)
ggplot(b, mapping = aes(x = chargeTimeHrs, y = kwhTotal)) +
  geom_point() +
  facet_wrap(~ facilityType) +
  labs(title = "pattern of chargeTimeHrs and kwhTotal depending on facilityType")
```



Do you see any differences in electric car charging pattern by type of facility?

Overall, there seems to be a general charging pattern in average time charged and the kwhTotal that each EV gets, but the Manufacturing facility seems to have a significantly higher range in kwhTotal that EVs get for the same amount of hours, possibly signifying a difference in quality in charging stations or type of EV. Similarly to the Manufacturing facility, the R&D facility has a wide spread in the kwhTotal and the chargeTimeHrs. Similarly, there might be a discrepancy with charging stations or EV models to explain the high kwhTotal each EV is getting; R&D workers seem to charge for longer times, possibly showing the long hours they work.

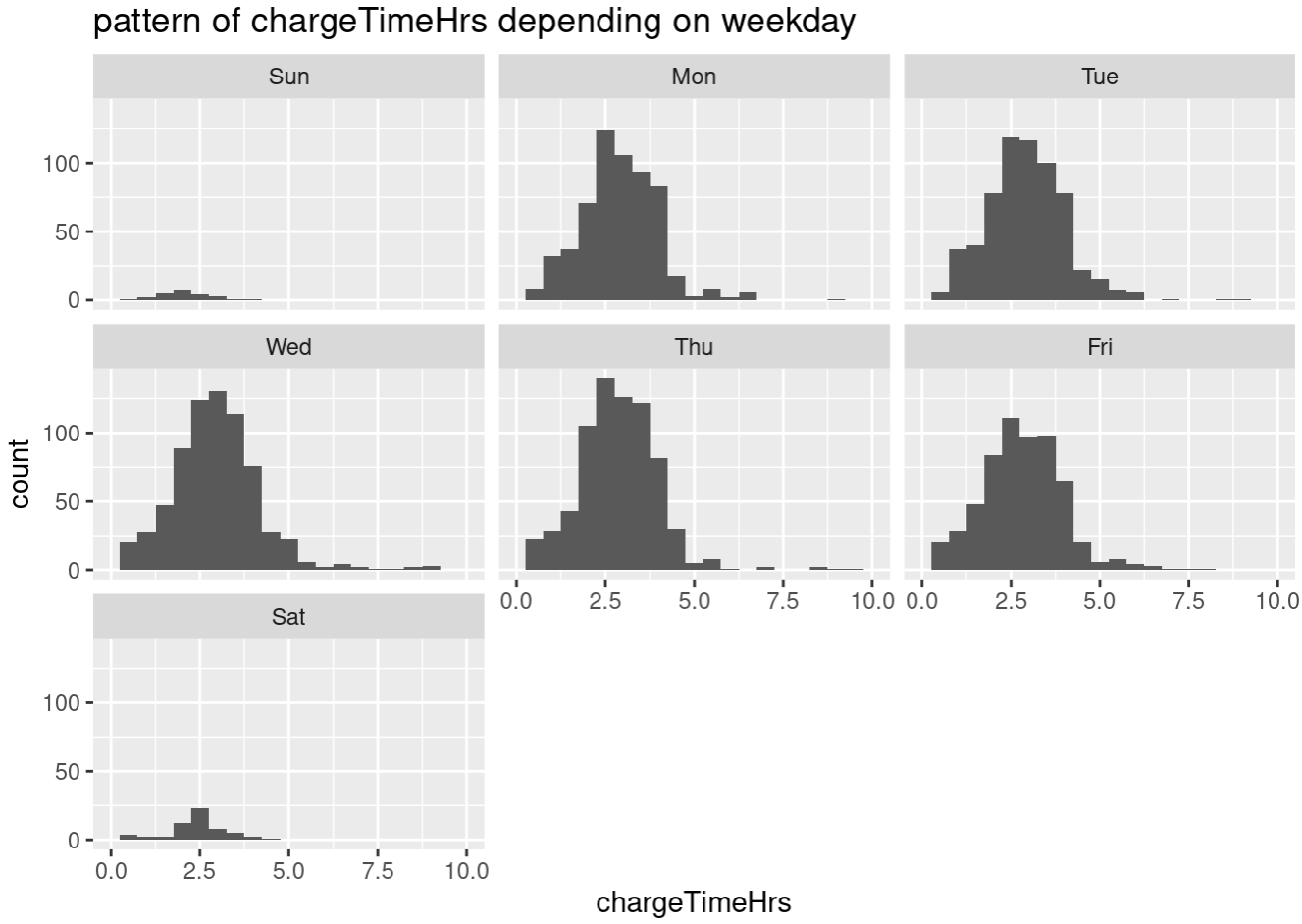
Question 8

Does the distribution of charging time vary by day of the week? Facet wrap a histogram to answer this question. Limit the range of the x axis from 0 to 10. Again, notice that all seven plots end up with the same x and y scales to make visual comparison valid.

```
ggplot(data = ev_cars, mapping = aes(x = chargeTimeHrs)) +
  geom_histogram(binwidth = 0.5) +
  facet_wrap(~ weekday) +
  xlim(0, 10) +
  labs(title = "pattern of chargeTimeHrs depending on weekday")
```

```
## Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 14 rows containing missing values (`geom_bar()`).
```



Do you see any differences in the distribution of electric car charging time by day of the week?

Yes, the most noticeable difference is that the weekends (Saturday and Sunday) have the least amount of charge times, mainly because employees typically do not work over the weekend, and therefore do not need to charge their EVs at their workplaces. On the weekdays, the distribution of data is overall very similar; typically unimodal with some smaller peaks on some days (Friday), which might be a result of employees working from home or leaving work earlier.

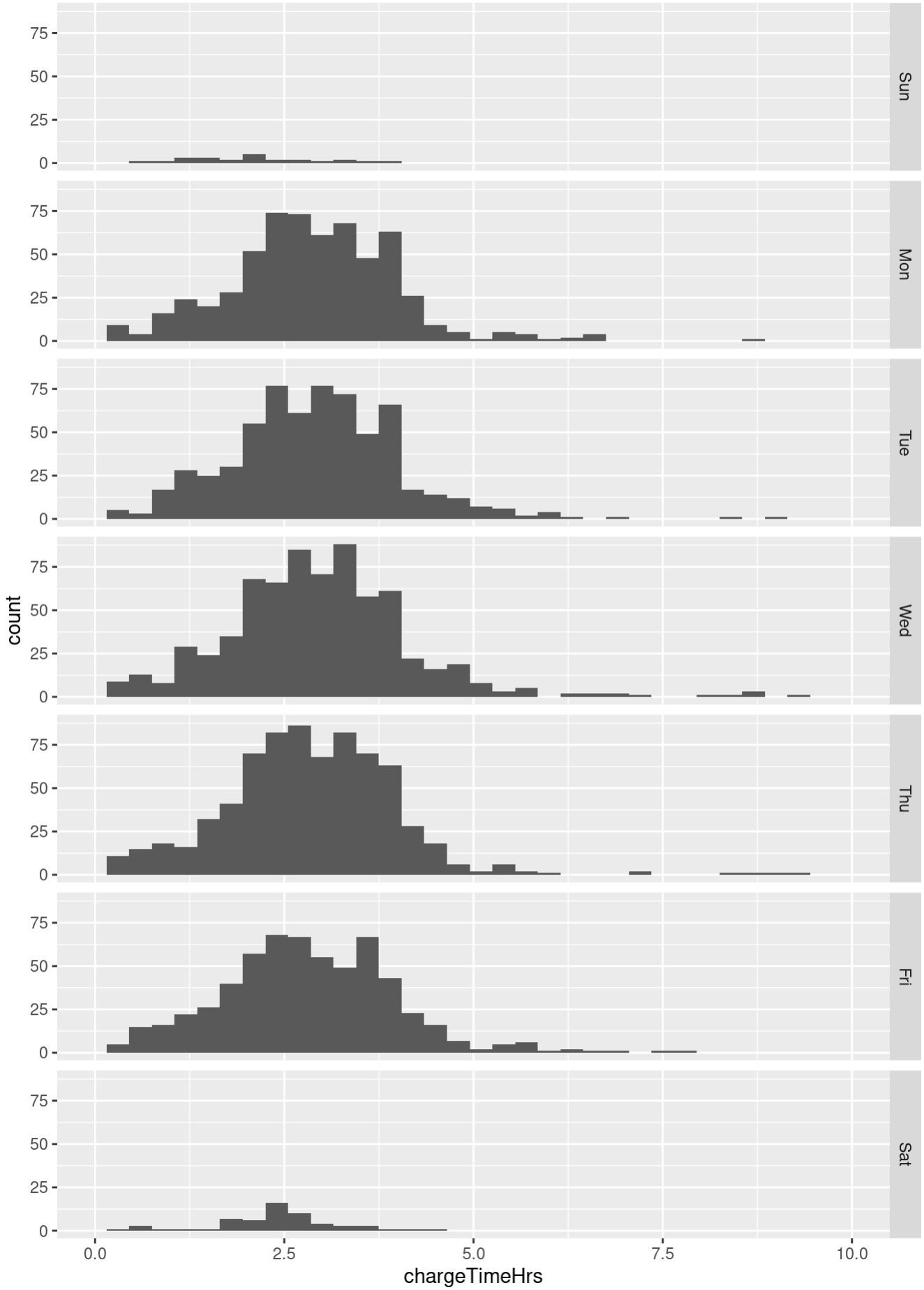
Question 9

To make the above analysis easier to interpret a single column facet grid maybe more appropriate. Implement a single column facet grid to answer the question above. Also set the figure height for the generated output to be 10 inches.

```
ggplot(data = ev_cars, mapping = aes(x = chargeTimeHrs)) +
  geom_histogram(binwidth = 0.3) +
  facet_grid(rows = vars(weekday)) +
  xlim(0, 10) +
  labs(title = "pattern of chargeTimeHrs depending on weekday-- faceted")
```

```
## Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 14 rows containing missing values (`geom_bar()`).
```


pattern of chargeTimeHrs depending on weekday-- faceted
count
chargeTimeHrs
Sun
Mon
Tue
Wed
Thu
Fri
Sat
0
25
50
75
0.0
2.5
5.0
7.5
10.0

Does the facet grid make the comparison of the data distributions easier to interpret? Why or why not?

Yes, the facet grid makes it easier to compare the data distribution throughout the week. This is because first of all, the facet wrap makes the graphs wrap into the next row, making it harder to see the side-by-side comparison as in the facet grid. Also, with the grid, we can see a more detailed distribution of the chargeTimeHrs, showing a better image of the skewed distribution and the amount of modes (unimodal/multimodal).

The Pledge

On my honor, I have neither given nor received any unacknowledged aid on this assignment.

Chaeyon Jang 12/04/2023