21235117_Assignment4

Diksha Srivastava

2022-03-28

Loading library

```
# loading libraries
library(readr)
library(dplyr)
library(ggplot2)
library(RColorBrewer)
library(lubridate)
library(scales)
library(ggalluvial)
library(forcats)
library(treemapify)
library(colorspace)
```

Reading file

```
# reading file
traffic_data <- read_csv(
  file = "E:/Users/Diksha/Desktop/NUIG/DV/Junction Turning Counts 2016 Outside.csv")
```

Task 1: Distribution of vehicles per 15 minute interval per vehicle type at the junction

Pre-processing

1. To be used in the graph, I have grouped the data by vehicle and time and found the sum of each vehicle at a particular time.
2. Since the distribution requires the full name of vehciles hence, I have used case_when() for multiple if-else conditions and created a new column in traffic_data named vehicleName.

```
# grouping data by vehicle and time to find the count of vehicles at each time.
grouped_data <- traffic_data %>%
  group_by(vehicle, TIME) %>%
  summarise(count = sum(count))

# renaming vehicles to their full name.
```

```
grouped_data$vehicleName <- case_when(
  grouped_data$vehicle == "PCL" ~ "Pedal Cycle",
  grouped_data$vehicle == "CAR" ~ "Car",
  grouped_data$vehicle == "TAXI" ~ "Taxi",
  grouped_data$vehicle == "MCL" ~ "Motorcycle",
  grouped_data$vehicle == "LGV" ~ "Light Goods Vehicle",
  grouped_data$vehicle == "OGV1" ~ "Ordinary Goods Vehicle 1",
  grouped_data$vehicle == "OGV2" ~ "Ordinary Goods Vehicle 2",
  grouped_data$vehicle == "CDB" ~ "City Direct Bus",
  grouped_data$vehicle == "BEB" ~ "Bus Eireann Bus",
  grouped_data$vehicle == "OB" ~ "Other Bus"
)

head(grouped_data)
```

```
## # A tibble: 6 x 4
## # Groups:   vehicle [1]
##   vehicle TIME                count vehicleName
##   <chr>   <dttm>              <dbl> <chr>
## 1 BEB     2016-11-23 07:00:00     0 Bus Eireann Bus
## 2 BEB     2016-11-23 07:15:00     2 Bus Eireann Bus
## 3 BEB     2016-11-23 07:30:00     0 Bus Eireann Bus
## 4 BEB     2016-11-23 07:45:00     0 Bus Eireann Bus
## 5 BEB     2016-11-23 08:00:00     0 Bus Eireann Bus
## 6 BEB     2016-11-23 08:15:00     1 Bus Eireann Bus
```

Plot description

For showing the distribution of vehicles, I have used strip plot along with boxplot because of the following reasons:

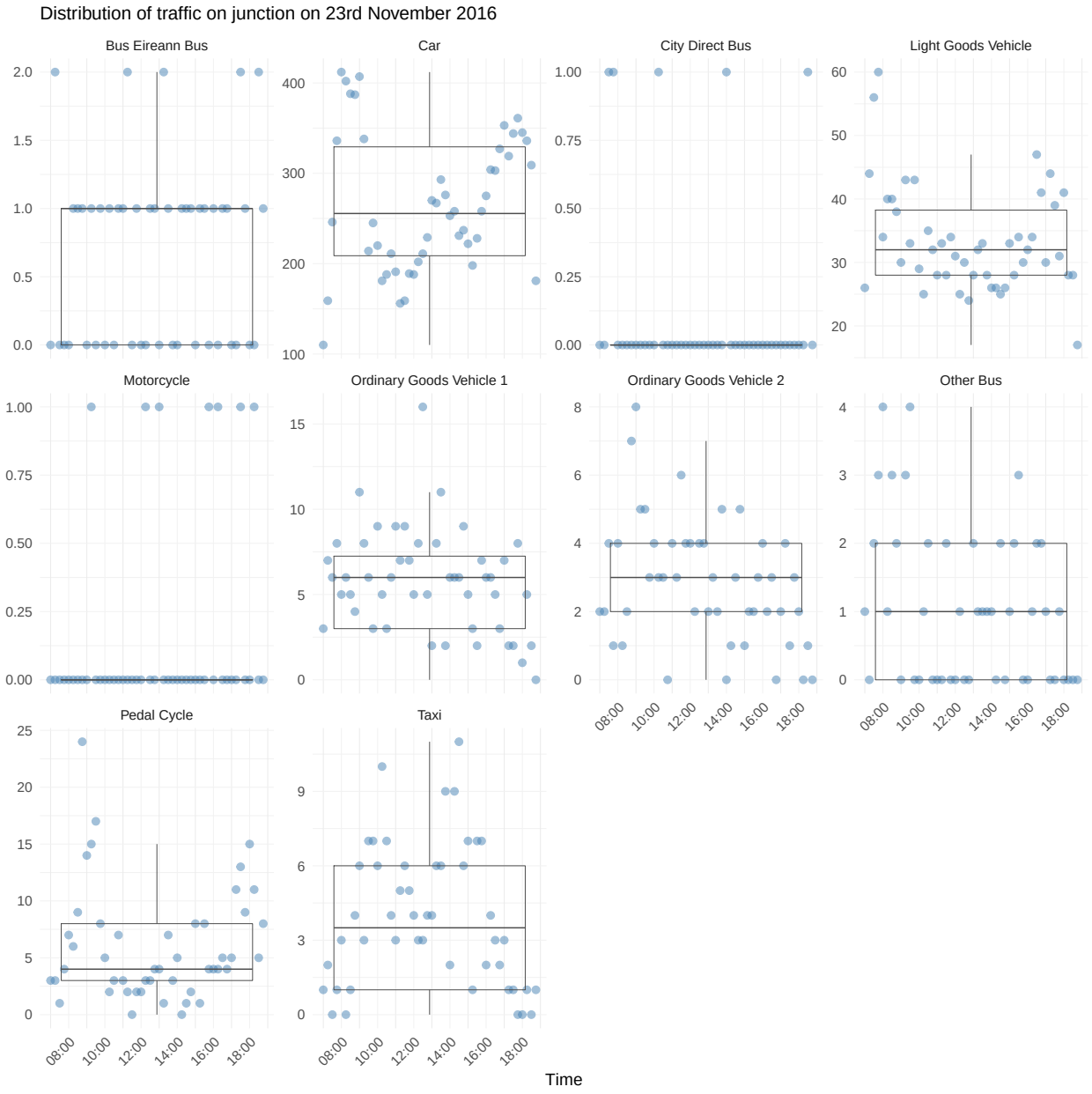
1. A strip plot will help in showing the exact value at each time interval for each vehicle. The x-axis could have time interval of 15 minutes throughout the day and y-axis could have the count of vehicle at that particular time. Since we don't need to see which turn the vehicle was taking thus we did group by vehicle to find the count of vehicle at that time.

2. Since I am plotting only one value for each time interval thus, I don't need to use jitter in my plot as there is no overlapping issue.

3. I have also used boxplot to show the summary statistics of the distribution over the time interval. Boxplot shows the minimum, 1st quartile, median, 3rd quartile, and maximum values for the data. 1st quartile represents where the 25% of the data lies. Similarly Median shows the 50% of data and 3rd quartile shows 75% of data lie till here. We can also idenitfy any outliers. However, when plotting with a strip plot outliers needs to be removed as individual points are already plotted and thus we can identify outliers easily. Detecting outliers would help us know that the volume of traffic is not high at a particular timestep. However, most of the volume of traffic will be where the boxplot is plotted.

```
ggplot(grouped_data, aes(x = TIME , y= count, group = vehicleName)) +
  # creating strip plot
  geom_point( size = 2, alpha = 0.5,
              color = "steelblue") +
  # creating boxplot.
```

```
  geom_boxplot(colour = "gray30",
               show.legend=F,
               # not showing the outliers
               outlier.shape = NA,
               varwidth = TRUE,
               alpha=0,
               size=0.2) +
  # showing the 2 hour interval on x-axis.
  scale_x_datetime(breaks = "2 hours",
                   name = "Time",
                   labels=date_format("%H:%M")) +
  # faceting based on vehicle name and free y scale so that each plot has different
  # y-axis.
  facet_wrap(~vehicleName, scales = "free_y") +
  ggtitle("Distribution of traffic on junction on 23rd November 2016") +
  theme_minimal() +
  theme (
    panel.grid.major.y =element_line(colour = "gray95", size =0.25),
    panel.grid.major.x =element_line(colour = "gray95", size =0.25),
    panel.grid.minor.y =element_line(colour = "gray95", size =0.15),
    # keeping the x-axis label at 45 degree
    axis.text.x = element_text(angle = 45,
                               vjust = 1,
                               hjust = 1),
    axis.title.y = element_blank(),
    legend.title = element_blank(),
    plot.title = element_text(size=12))
```



Visual elements

Length: Each individual point in a strip plot is represented by dot. If the dot is higher or has more length from x-axis then it means it has higher value that means it contributed more to traffic at that time and vice-versa.

Facet: I have facetted the plot according to category of vehicle so that overlapping is less and since the range of values of count varied widely as cars had values till 400 and other had in the range of 1-5 thus plotting them on different scale would help in analysing the plot.

Interpretation

1. It is clear that the cars contribute the most in the traffic as it's value lies in the range 100-400. Also the boxplot indicates that most of the counts over the time inetrval were in range 200-350. 25% of the data had count below 210 and 75% of data had count below 330. The minimum count was 100 and the maximum count was approx. 400.
2. Next, light goods vehicle had contributed more to traffic as compared to others except cars. Its value ranged from 20-60. Most of the counts were in range 28 to 38 and minimum value was 16 and maximum was 47. There were some outliers that had value 60 in early morning.
3. Pedal cycle's had distribution in the range 0 to 25. However, the distribution is concentrated between 3 to 7. Minimum value is 0 and maximum is 15 and there is one outlier in morning as well.
4. Only 2-4 ordinal goods vehicle 2 contributed to the traffic at junction with a minimum value of 0 at some times and maximum of 7 at some times.
5. Bus Eirean Bus, Other bus, and Taxi had a wide boxplot indicating that their counts were distributed throughout their scale. Bus Eireann and Other buses had count 0 25% of the times. Other bus had count 1 or below 75% of times.
6. Motor cycle and city direct bus had either count 0 or 1 indicating that it contributed very less to traffic.

Task 2: Proportions of traffic coming from D to A, B, and C.

Pre-processing

1. Since for this task, we only required data that had turns from D to A, B, and C, thus, I filtered out the data having turn as DA, DB, and DC.
2. We had to divide the time in intervals, so for that I first extracted the time from data-time column TIME. I used format() method for the same and extracted the time in hours-minutes format using %H:%M. I updated the TIME column with only the time values that was extracted.
3. I created a new column in the filtered_data that had the time interval. I used case_when() to apply multiple if-else conditions to divide the time in intervals like 07:00 to 09:30 will be in an interval Early morning. I included 07:00 in the early morning and excluded 09:30. Thus, early morning will have intervals from 07:00 AM to 09:29 AM.
4. The intervals created were converted to factor so that it can be used further in the plot.
5. On the resultant data, I grouped them based on turn and intervals obtained and calculated their sum so that we can identify their proportion while making the plot.

```
# filtering out traafic from D to A, B, and C.
filtered_data <- traffic_data %>% filter(turn == 'DA' | turn == 'DB' | turn == 'DC')
# converting the date-time to time format.
filtered_data$TIME <- format(filtered_data$TIME, format="%H:%M")
# assigning time slots to intervals and converting them to factor.
filtered_data <- filtered_data %>% mutate(interval = case_when(
  filtered_data$TIME >= '07:00' & filtered_data$TIME < '09:30' ~ "Early morning",
  filtered_data$TIME >= '09:30' & filtered_data$TIME < '12:00' ~ "Late morning",
  filtered_data$TIME >= '12:00' & filtered_data$TIME < '14:30' ~ "Afternoon",
  filtered_data$TIME >= '14:30' & filtered_data$TIME < '17:00' ~ "Late afternoon",
  filtered_data$TIME >= '17:00' & filtered_data$TIME < '19:00' ~ "Evening")) %>%
  mutate(interval=factor(interval,
```

```
                          levels = c("Early morning",
                                     "Late morning",
                                     "Afternoon",
                                     "Late afternoon",
                                     "Evening")))

# grouping the data based on DA, DB, and DC and calculating the traffic sum using
# summarize.
filtered_data <- filtered_data %>%
  group_by(turn, interval) %>%
  summarize(count=sum(count))

head(filtered_data)
```

```
## # A tibble: 6 x 3
## # Groups:   turn [2]
##   turn  interval       count
##   <chr> <fct>          <dbl>
## 1 DA    Early morning     69
## 2 DA    Late morning      43
## 3 DA    Afternoon         48
## 4 DA    Late afternoon    19
## 5 DA    Evening           12
## 6 DB    Early morning    196
```

Plot description

For representing the proportion of traffic going from D to A, B, and C, I have chosen Alluvial plot.

Reasons for choosing alluvial plot:

1. An alluvial plot helps in representing proportional relationship between variables. It also helps in determining the flow from one categorical variable to another. In our case since, first we needed to show how the traffic coming from D divides into A, B, and C. Thus, the first axis of the alluvial plot represents the proportion of traffic that goes to A (represented by DA), to B (DB), to C (DC). The aesthetic used here is length such that if the length of category is more then it means that the traffic was in higher proportion in that category.

2. Next, we needed to show how the traffic for each category (DA, DB, and DC) is distributed over different times of day. The second axis show shows different times of day. The flow from the first axis to the second axis shows how the traffic of each category DA, DB, and DC is distributed over different times of day. We are able to identify the flow for different categories as we have used the color aesthetic, such that different categories are represented by different color.

3. It is easy to follow the flow from DA, DB, and DC to their corresponding time interval i.e., early morning, afternoon etc.

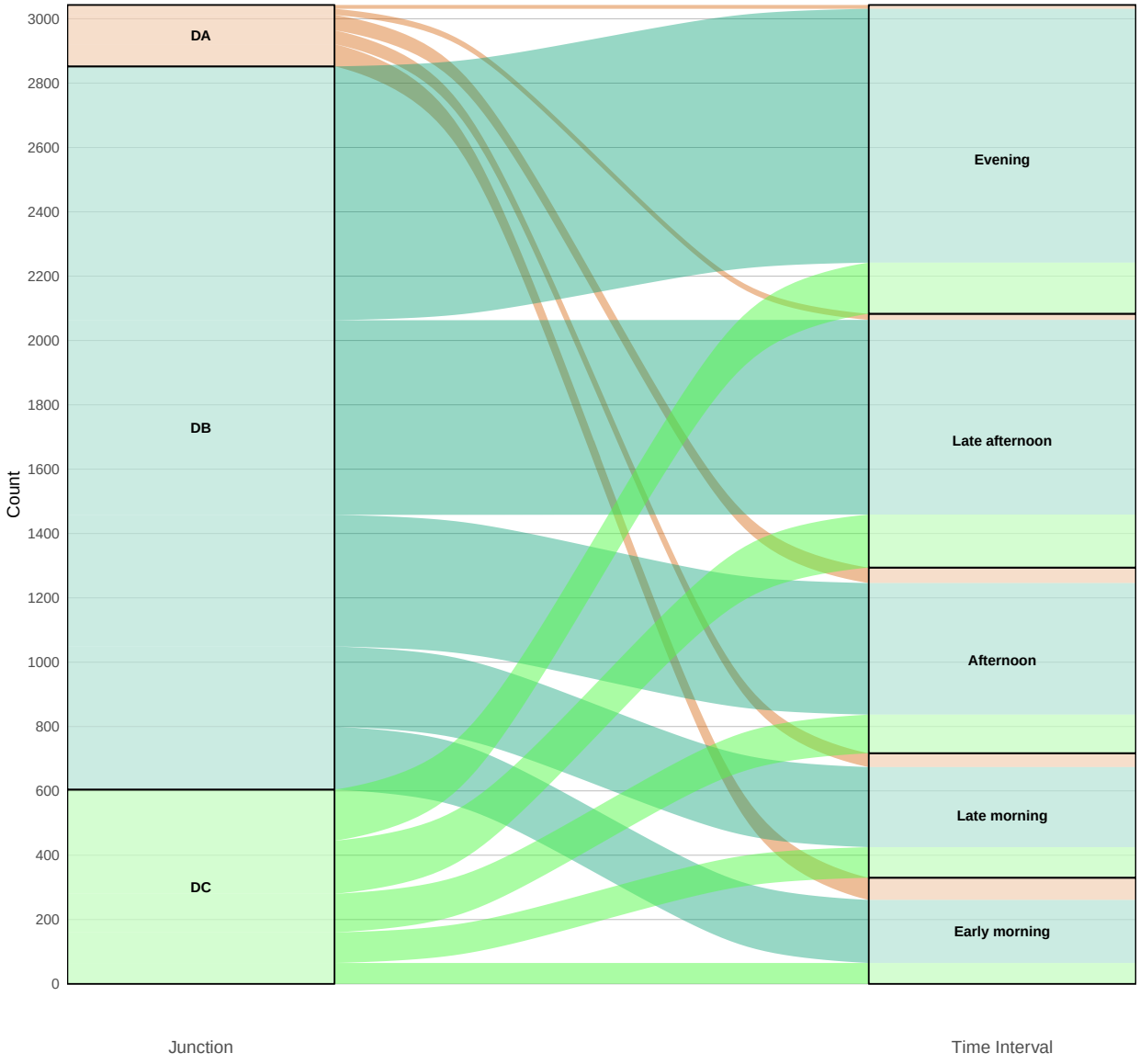
4. An alluvial plot is preferred over parallel plots because we can order the axis values. For example, we can order the time intervals like early morning, late morning, afternoon, late afternoon, and evening. In case of parallel plots these would have come in an unordered way. Also the alpha values can be changed so it is easier to follow the path during overlapping as well.

5. An alluvial plot unlike parallel plot splits the data from the starting axis. Example, we can see the color is assigned from the first axis based on 3 categories DA (red), DB (blue), and DC (green). Also the color runs till the end axis i.e., we can see the DA proportion in early morning and so on.

```r
# to plot two categorical axis, turn their respective columns to factor.
# ordering the time slots in reverse factor.
filtered_data$interval <- fct_rev(filtered_data$interval)

ggplot(data = filtered_data,
       # taking axis 1 as DA, DB, DC.
       aes(axis1 = turn,
           # taking axis 2 as Early morning and so on.
           axis2 = interval,
           # y axis is the count grouped by turn and intervals so that we have
           # count of DA for each interval and so on.
           y = count)) +
  # setting the name of x-axis.
  scale_x_discrete(limits = c("Junction", "Time Interval"),
                   expand = c(.01, .05),) +
  # setting the name and breaks of y-axis.
  scale_y_continuous(name = "Count",
                     breaks = seq(0, 3100, by = 200)) +
  # providing colors for categories DA, DB, DC.
  scale_fill_manual(
    values = c("DA" = "#D55E00D0", "DB" = "#009E73D0", "DC" = "#54FA47"),
    guide = "none") +
  # plotting allucium plot based on turns DA, DB, DC.
  geom_alluvium(aes(fill = turn)) +
  geom_stratum(fill="white", alpha=0.5) +
  ggtitle("Proportions of traffic from D to A, B, and C at different times of day") +
  # setiing label using after_stat
  geom_text(stat = "stratum",
            aes(label = after_stat(stratum)),
            size=3, fontface="bold") +
  theme_minimal() +
  theme(panel.grid.minor.y= element_blank(),
        panel.grid.major.y= element_line(size=0.1, colour = "grey"),
        panel.grid.major.x= element_blank(),
        axis.text.x = element_text(size=11))
```

Proportions of traffic from D to A, B, and C at different times of day



Visual Elements

1. Color: to show category if the flow is from DA, DB, or DC.
2. Length: Shows proportion of traffic from D to A, B, and C.

Interpretation

1. It is clear that most of traffic on the junction is because of the vehicles moving from D to B. More than 50 percent of traffic is because of vehicles moving from D to B. Furthermore, the vehicles moving from D to C contributes next (approximately 25 percent) to traffic at junction and vehicles moving from D to A contributes least approximately 10 percent of the traffic.

2. Vehicles moving from junction D to B and D to C contributes most to the traffic in the evening (17:00 to 19:00). The proportion of traffic from these two paths is still high in the late afternoons as well but slightly less than the evening. Similarly, the traffic contributed by vehicles on these two paths decreases in afternoon, further reduces in late morning. The traffic contributed in early morning is least as compared to other time slots.

3. For vehicles following path D to A, most of the traffic is in early morning, which further reduces throughout the day and is least in the evening.

4. Overall, the proportion of traffic was due to vehicles moving from D to B then from D to C and least by D to A. Below are some approximate proportions of traffic for different categories for some time intervals:

DA: 10% traffic

DB: 65% traffic

DC: 25% traffic

Evening: DA: 2%

DB: 70%

DC: 28%

Late Afternoon: DA: 5%

DB: 67%

DC: 28%

Early morning: DA: 25%

DB: 55%

DC: 20%

Note: These are only approximate values based on what can be percieved from the graph.

Task 3: Volume of vehicles per vehicle type at each timestep.

Pre-processing

No pre-processing required as the data had already been prepared in 1st question. Using the same data to do the plot.

Plot Description

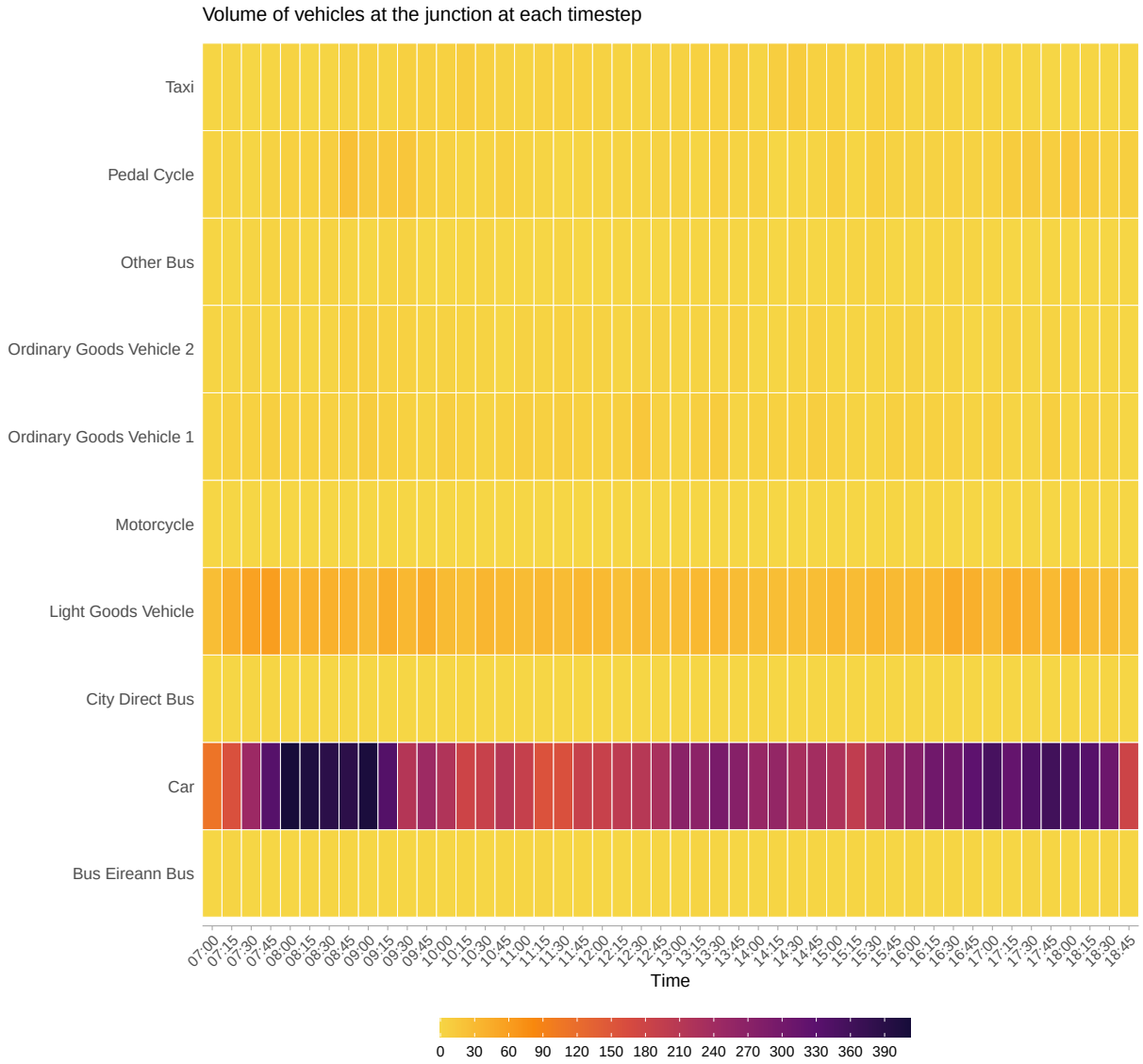
For this task, I chose to use HeatMap because of the following reasons:

1. We had 3 values which we wanted to show in a graph. First, each timestep from 7:00AM to 7:00PM, second, all the categories of vehicles and third count of the vehicles at each timestep. We didn't have to show the actual count of traffic but just the patterns which represents the volume of traffic at the junction at each timestep.

2. HeatMap is good for showing any kind of patterns in the data and to represent multiple variables effectively. It helps in detecting variance across multiple variables, if variables are similar to each other, and if there is any correlation between them.

3. All rows are represents one category each i.e., Vehicle type and all columns represent one category each i.e., 15 minute interval timestamp. The cells represent the count of traffic for that category at a particular timestep. It is represented by colour variations. Thus, we create a time series heat map to show the volume of traffic at each timestep.

```
# plotting heat map with timesteps on x-axis and vehicle names on y-axis.
# it is filled with count of traffic.
ggplot(grouped_data, aes(x= TIME, y=vehicleName, fill = count)) +
  geom_tile(colour = "white") +
  # removing y-axis label
  scale_y_discrete(name = NULL) +
  # setting the format of time in hours and minutes.
  scale_x_datetime(name = "Time",
                   expand = c(0,0),
                   date_breaks = "15 mins",
                   date_labels = "%H:%M" ) +
  # using viridis to color the traffic variations in heatmap.
  # choosing the color variation.
  scale_fill_viridis_c(option = "B",
                       # setting the range of color scales to be used.
                       begin = 0.1,
                       end = 0.9,
                       direction = -1,
                       # setting the scale name blank.
                       name = "",
                       # setting the height and width of scale bar at the bottom of graph.
                       guide = guide_colourbar(direction = "horizontal",
                                               barwidth = 20,
                                               barheight = 0.8),
                       # setting the scale of the graph.
                       breaks = seq(0, 450, by= 30)) +
  ggtitle("Volume of vehicles at the junction at each timestep") +
  theme(axis.text.y = element_text(size=10),
        axis.ticks.x = element_line(size=0.3, colour = "darkgrey"),
        axis.line.x = element_line(size=0.3, colour = "darkgrey"),
        axis.ticks.y = element_blank(),
        axis.line.y = element_blank(),
        # setting the background blank.
        panel.background = element_blank(),
        panel.grid = element_blank(),
        # rotating the x-axis label by 45 degree.
        axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   hjust = 1),
        plot.margin = unit(c(0.5, 0.5, 2, 0.5), "cm"),
        plot.title = element_text(size=12),
        # moving the scale to the bottom of the graph.
        legend.position = "bottom")
```



Visual Elements

color: used to show the volume of traffic. A dark color represents more traffic while a light color represents less traffic. The raw value of the traffic can be approximated from the scale at the bottom which is divided into several breaks.

Interpretation

1. It is clearly visible that the volume of traffic contributed by Car is highest as compared to other vehicles. The range of car count throughout the 12 hour period lies between 90 to 420 while for the other vehicles most of the values lie below this range. It is also visible that the volume of traffic contributed by car is highest in the morning around 8:00AM to 9:00AM with values above 360. This

count decreases in the late morning from 9:30AM to 12:30 and lies in the range 150 to 240. In the afternoon, the count increases slightly between 240 to 270. Later in the evening the count is high but not higher than the morning. Between 15:45 to 18:15 the value lies in the range 270 to 390. This may be due to many people going to office in the morning, and returning in the evenin and using car as a means of transport.

2. Bus Eireann Bus, City Direct Bus, Other Bus, and Motorcycle do not show any variation throughout the 12 hour period. Their values lies consistently in the range 0 to 30.

3. Light Goods Vehicle contributed to traffic a little higher than other vehicles except car. The highest value was during the early morning from 7:15 to 7:45 and around 60 vehicles contributed to the traffic. Throughout the day around 30 to 60 vehicles contributed to the traffic. In the evening the value increased again close to 60.

4. A very slight variation in volume of traffic can be seen in Ordinary Goods Vehicle 1, Pedal cycle, and Taxi. OGV1 values lied in the range 0 to 30 throughout the period. Most of the values were in the lower scale of 0-30 while only 2/3 values can be seen closer to 30 at around 12:30 in the afternoon. For pedal cycle as well, all the values lies in the range 0 to 30, however, between 8:30AM to 9:15AM, the cycles were near to 30. For Taxi as well, all the values lied in the range 0 to 30, however, at around 14:45 the taxi were slightly higher than other times of the day.

Task 4: Proportion of categories of Vehicle and their sub-categories

Pre-processing

1. For this task, I need the count of each vehicle contributing to the traffic. Thus, I grouped the traffic data by vehicle and found the sum of each vehicle for the 12 hour period.

2. I also created a new column which would assign and store the category for each vehicle. I used case_when for multiple if-else conditions and stored the categories of each vehicle in category column.

```
# finding the count of traffic for each vehicle.
tree_data <- traffic_data %>%
  group_by(vehicle) %>%
  summarise(count = sum(count))

# assigning categories to each vehicle.
tree_data$category <- case_when(
  tree_data$vehicle == "PCL" ~ "Two-wheel vehicles",
  tree_data$vehicle == "CAR" ~ "Cars",
  tree_data$vehicle == "TAXI" ~ "Cars",
  tree_data$vehicle == "MCL" ~ "Two-wheel vehicles",
  tree_data$vehicle == "LGV" ~ "Goods Vehicle",
  tree_data$vehicle == "OGV1" ~ "Goods Vehicle",
  tree_data$vehicle == "OGV2" ~ "Goods Vehicle",
  tree_data$vehicle == "CDB" ~ "Buses and public transport",
  tree_data$vehicle == "BEB" ~ "Buses and public transport",
  tree_data$vehicle == "OB" ~ "Buses and public transport"
)

head(tree_data)
```

```
## # A tibble: 6 x 3
##   vehicle count category
```

```
##   <chr>   <dbl> <chr>
## 1 BEB        32 Buses and public transport
## 2 CAR     12717 Cars
## 3 CDB         5 Buses and public transport
## 4 LGV      1602 Goods Vehicle
## 5 MCL         7 Two-wheel vehicles
## 6 OB         50 Buses and public transport
```

Plot Description

For this task, I have chosen TreeMap because of the following reasons:

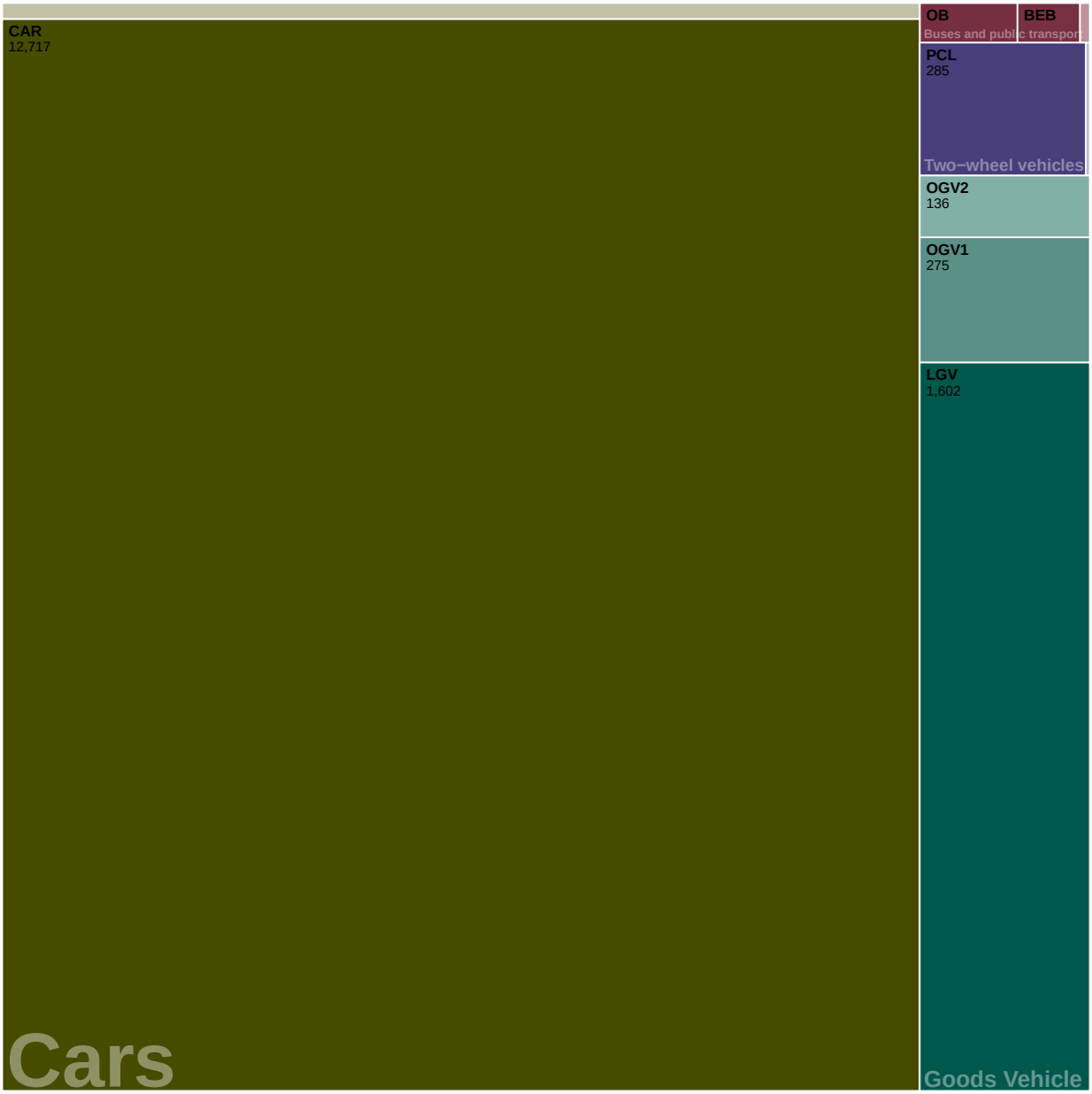
1. According to the question, we have to show proportion of the categories of vehicles and sub-categories of vehicles. Thus, this is a hierarchical data. TreeMap is used to represent heirarchy of data. It would represent quantities of each category and sub-category according to area size.
2. The main categories like Cars, Buses and public transport etc. will be assigned an area on the plot according to the sum of quantities in its sub-category. Each sub-category like BEB, CDB etc. would be present inside the main category and would be assigned an area according to the count of the vehicle on the junction. Thus, each category is proportional and each sub-category is also proportional.
3. I chose a separate color scale for each main category instead of one single color scale. It is because since most of the vehicles had quite less value and almost in the similar range, they were given the almost similar shade and it was difficult to identify any difference in their proportions.

```
# used to find the number of main categories so that many colours could be generated.
n <- length(unique(tree_data$category))

# calculating the colors for each vehicle.
tree_data_df <- tree_data %>%
  # creating index for each category.
  mutate(index = as.numeric(factor(category))- 1) %>%
  group_by(index) %>%
  mutate(
    max_count = max(count),
    # assigning hcl pallete colors to each vehicle.
    colour = gradient_n_pal(
      sequential_hcl( 4,
        h = 360 * index[1]/n,
        c = c(45, 20),
        l = c(30, 80),
        power = 0.01)
      )(1- (count/max_count))
    )


ggplot(data = tree_data_df ,
       # using the count of each vehicle to assign area.
       aes(area = count ,
           fill = colour,
           # assigning main categories as subgroup.
           subgroup = category)) +
  geom_treemap(colour = "white",
               size = 0.5*.pt,
```

```
            alpha = NA)  +
# assigning label to each area.
geom_treemap_text(aes(label = vehicle),
                  colour = "black" ,
                  size =10,
                  place = "topleft",
                  fontface = "bold",
                  padding.x = grid::unit(1.5, "mm"),
                  padding.y = grid::unit(1.5, "mm")) +
# assigning the count of vehicle to each area.
geom_treemap_text(aes(label = format(count, nsmall=0, big.mark=",",trim=TRUE)),
                  color = "black",
                  size = 9,
                  place = "topleft",
                  min.size = 3,
                  padding.x = grid::unit(1.5, "mm"),
                  padding.y = grid::unit(15, "points")) +
# assigning border and text to main category.
geom_treemap_subgroup_border(colour = "white", size =0.5) +
geom_treemap_subgroup_text(grow = FALSE,
                           colour = "#FAFAFA",
                           size = 50,
                           place ="bottomleft",
                           fontface = "bold",
                           alpha = 0.4) +
# since data is already scaled to colour value, thus it will be used to define
# the aesthetics instead of ggplot using its own scale.
scale_fill_identity()+
guides(colour = "none", fill = "none")
```



Visual Elements

Area: Representing the count of traffic contributed by each vehicle. Color: Representing the parent categories Cars, Goods Vehicle, Two-wheel vehicles, Bus and public transport. Sequential Color scales: Representing vehicles inside each category car, taxi, LGV, OGV1, OGV2, PCL, MCL, OB, and BEB. Scale: Every area except the ones having low proportion in treemap is given a value which represents the actual count of that vehicle contributing towards traffic in 12 hour period.

Interpretation

1. We can see that there are 4 categories namely Cars, Goods Vehicle, Two-wheel vehicles, Buses and public transport. The main category Cars is in higher proportion as compared to others as occupies

the maximum area of the plot. Goods Vehicle is in less proportion as compared to cars and occupies area of 2013/15297. Two wheel vehicle is in less proportion than Goods vehicle and occupies an area of 292/15297. Buses and public transport are in very less proportion as compared to any other categories. Thus, it shows that most of the traffic is contributed by Cars and least contributed by Buses and public transport.

2. In the Cars category, Car occupies maximum area with 12717 count and Taxi occupies only 188 area i.e., proportion of taxi contributing to traffic is very less than Cars. In the plot, the label taxi is in so less amount that its label is not visible.

3. In Goods vehicle category, Light Goods vehicle occupies maximum area in this category with 1602 count out of 2013. The proportion Ordinary goods vehicle 1 is higher than Ordinary goods vehicle 2. This is obvious from the color scale. The color of OGV1 is same as OGV2 but there is difference in scale of both. OGV1 is of darker color than OGV2.

4. In Two-wheel vehicle, PCL has a larger proportion than MCL with value 285. MCL occupies a very small area so it is hardly visible. PCL is of much darker shade than MCL thus shows that it has a higher count and contributes more to traffic.

5. In buses category, Other buses occupies a higher proportion as compared to CEB, and BEB. CEB is not much visible as its value is quite less as compared to other areas.