# Analyzing Road Usage Data

Sunday April 02, 2023

## Contents

# 1 Background

In this analysis, I use data regarding road usage to explore factors that have a relationship with the extent of road use. I start by summarising the results and stating the objectives/questions of the study. I then explore the data. In the analysis section, I delve deeper into finding answers to my questions. The analysis then concludes.

# 2 Summary of Results

1. The distribution of road length depends on the road type.
2. The usage of a road has a significant relationship with the road type.
3. Longer roads have, on average, more vehicles than shorter roads.
4. National roads are, on average, longer than other road types.
5. The usage of a road has a direct relationship with the number of lanes, with 3-lane and 4-lane roads the most used.
6. Which road types are dominated by trucks?

# 3 Questions

The analysis seeks answers to the following questions.

- What can be said about the distribution of road length?
- What can be said about the distribution of road length for departmental roads?
- To what extent does the number of vehicles using a road depend on the length of that road?
- To what extent does the number of vehicles using a road depend on the length of that road for the different road types?
- Are departmental roads shorter on average than national roads?
- Is there a difference in length depending on the number of lanes?

# 4 Data

I import the data from the Ms Excel file and check for duplicates and missing values.

```
roads <- import("COPY_FR_donnees_routieres_simulees_avecEbaucheAnalyse.xlsx") %>%
  janitor::clean_names() %>%
  filter(!duplicated(.)) %>%
  mutate(type_route = factor(type_route))

roads %>% head() %>%
  kbl(booktabs = TRUE, caption = "Data Preview") %>%
  kable_classic(full_width = TRUE, latex_options = "hold_position")
```

The data contains 510 observations of 8 variables. Table 1 describes the variables.

Next, I visualise the data for missing values. The visualisation shows that there are no missing values.

```
roads %>%
  Amelia::missmap()
```

Finally, I summarise the data.

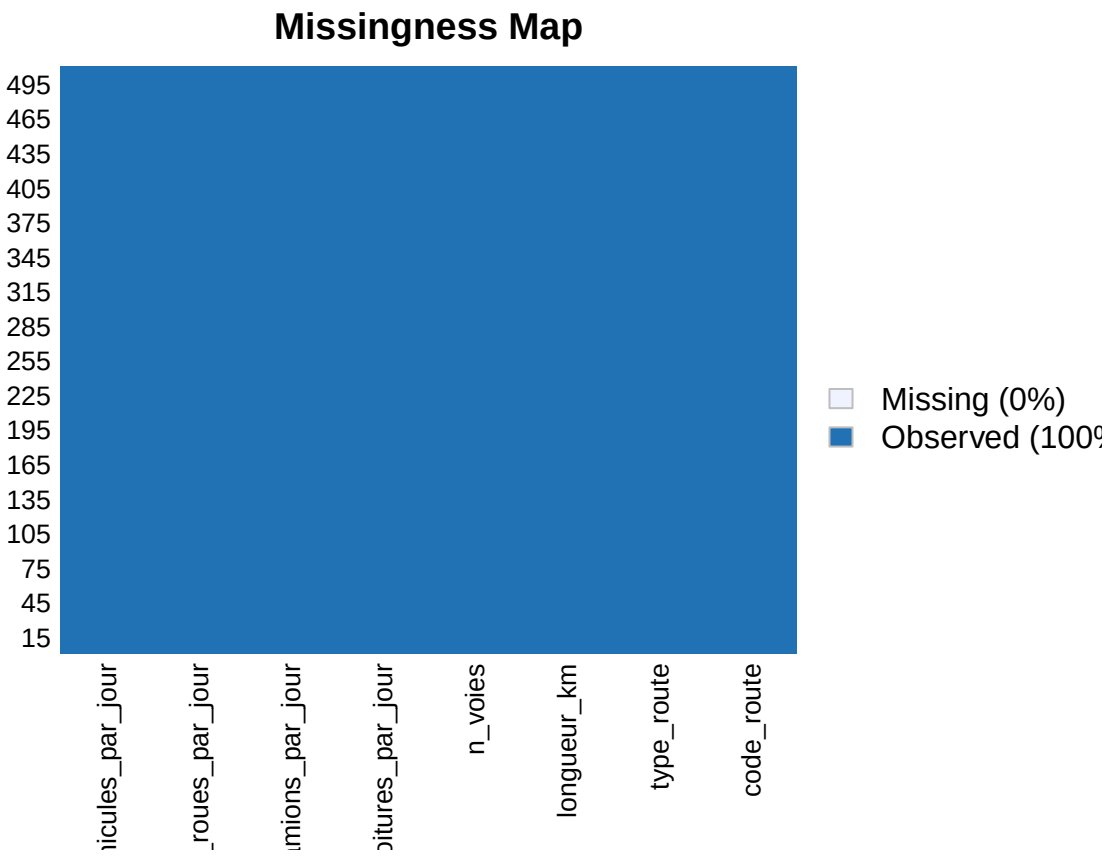



Figure 1: Missingness Map

Table 1: Data Preview

| code_route | type_route | longueur_km | n_voies | n_voitures_par_jour | n_camions_par_jour | n_deux_roues_par_jour | n_vehicules_par_jour |
|---|---|---|---|---|---|---|---|
| A1 | autoroute | 119 | 3 | 2026 | 331 | 656 | 3013 |
| A2 | autoroute | 507 | 3 | 2169 | 440 | 522 | 3131 |
| A3 | autoroute | 362 | 4 | 2301 | 465 | 448 | 3214 |
| A4 | autoroute | 538 | 3 | 1841 | 463 | 596 | 2900 |
| A5 | autoroute | 150 | 3 | 2710 | 629 | 447 | 3786 |
| A6 | autoroute | 454 | 3 | 2642 | 507 | 555 | 3704 |

Table 2: Variables Description

| Variable | Description |
|---|---|
| code_route | A code representing the route, e.g. A1. |
| type_route | A classitication of route type like autorute, nationale. |
| longueur_km | The length of the route in Kiliometers. |
| n_voies | Number of channels or lanes in the route. |
| n_voitures_par_jour | Number of cars using the route per day. |
| n_camions_par_jour | Number of trucks using the route per day. |
| n_deux_roues_par_jour | Number of 2 wheel vehicles using the route per day. |
| n_vehicules_par_jour | The sum of vehicles of all types on the route per day. |

```
## Summary of numeric variables
roads %>%
  select(where(is.numeric)) %>%
  skimr::skim_without_charts() %>%
  select(-skim_type, -n_missing, -complete_rate) %>%
  set_names(c("Variable", "Mean", "SD", "Min", "Q1", "Median", "Q3", "Max")) %>%
  kbl(booktabs = TRUE, caption = "Summary Statistics") %>%
  kable_classic(full_width = FALSE, latex_options = "hold_position")
```

Table 3: Summary Statistics

| Variable | Mean | SD | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|
| longueur_km | 52.2 | 97.21 | 0 | 5.3 | 9.5 | 52 | 576 |
| n_voies | 1.2 | 0.55 | 1 | 1.0 | 1.0 | 1 | 4 |
| n_voitures_par_jour | 241.8 | 386.90 | 33 | 119.0 | 149.0 | 184 | 2710 |
| n_camions_par_jour | 44.0 | 99.86 | 5 | 10.0 | 13.0 | 32 | 632 |
| n_deux_roues_par_jour | 73.6 | 95.73 | 6 | 38.0 | 50.0 | 73 | 658 |
| n_vehicules_par_jour | 359.4 | 574.75 | 88 | 180.0 | 216.0 | 265 | 3786 |

# 5 Analysis

For the analysis, I start by creating a pairs plot for all the variables except for code route that has too many levels.

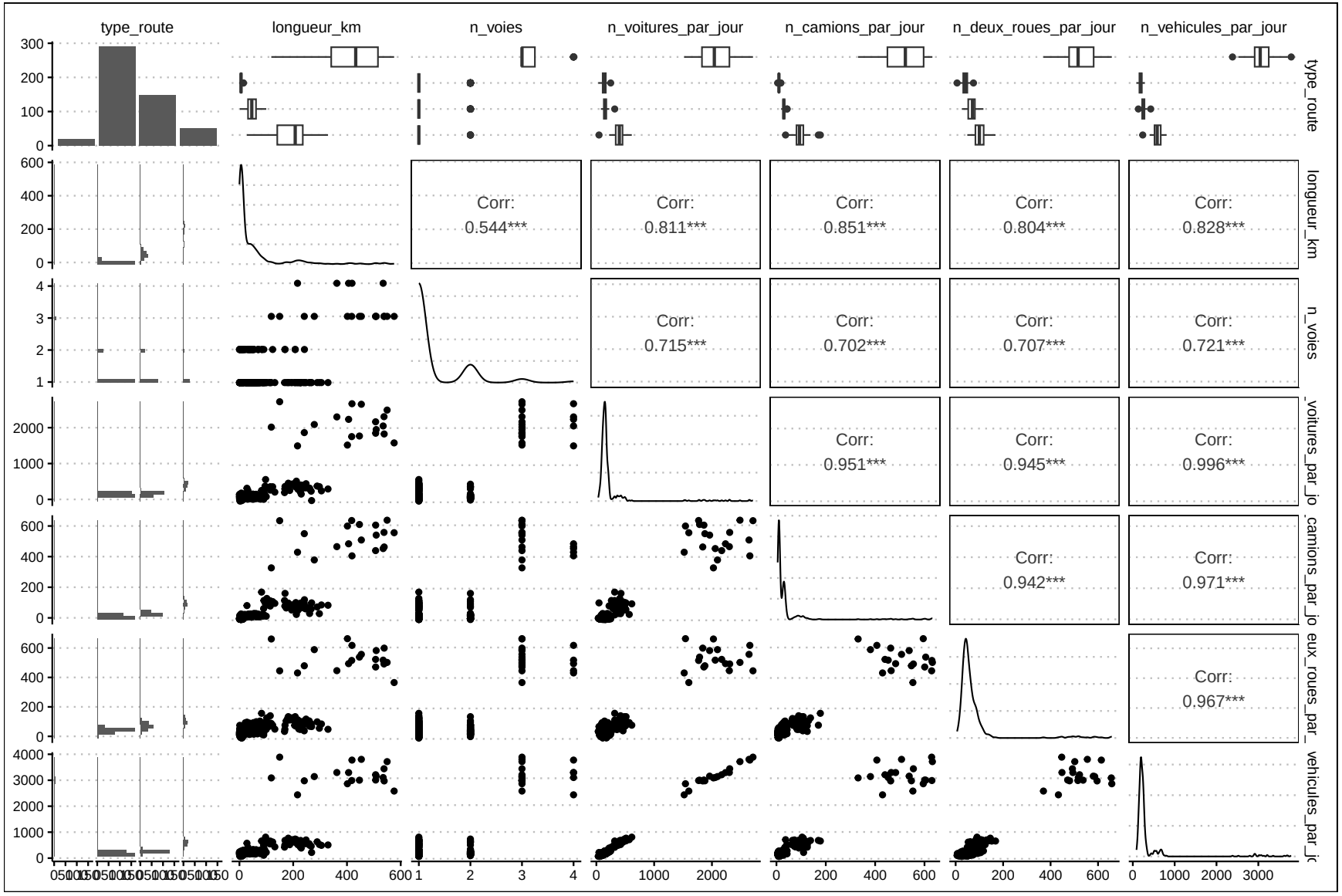



Figure 2: Pairs Plot for the Variables

Figure 1 shows strong correlation between the variables. This observation means that, for example, when there are more cars in the route, it is also highly probable that there will be more trucks. We now delve into our questions.

## 5.1 What can be said about the distribution of road length?

I use a histogram to visualise road length. A histogram (together with density plots) are useful for visualizing the distribution of a single quantitative variable. Figure 2 shows that most roads are less than 100 km long, with another small up tick around 200 kilometres.

```
roads %>%
  ggplot(mapping = aes(x = longueur_km)) +
  geom_histogram() +
  labs(x = "Road Length in KMs", y = "Count", title = "Distribution of Road length")
```

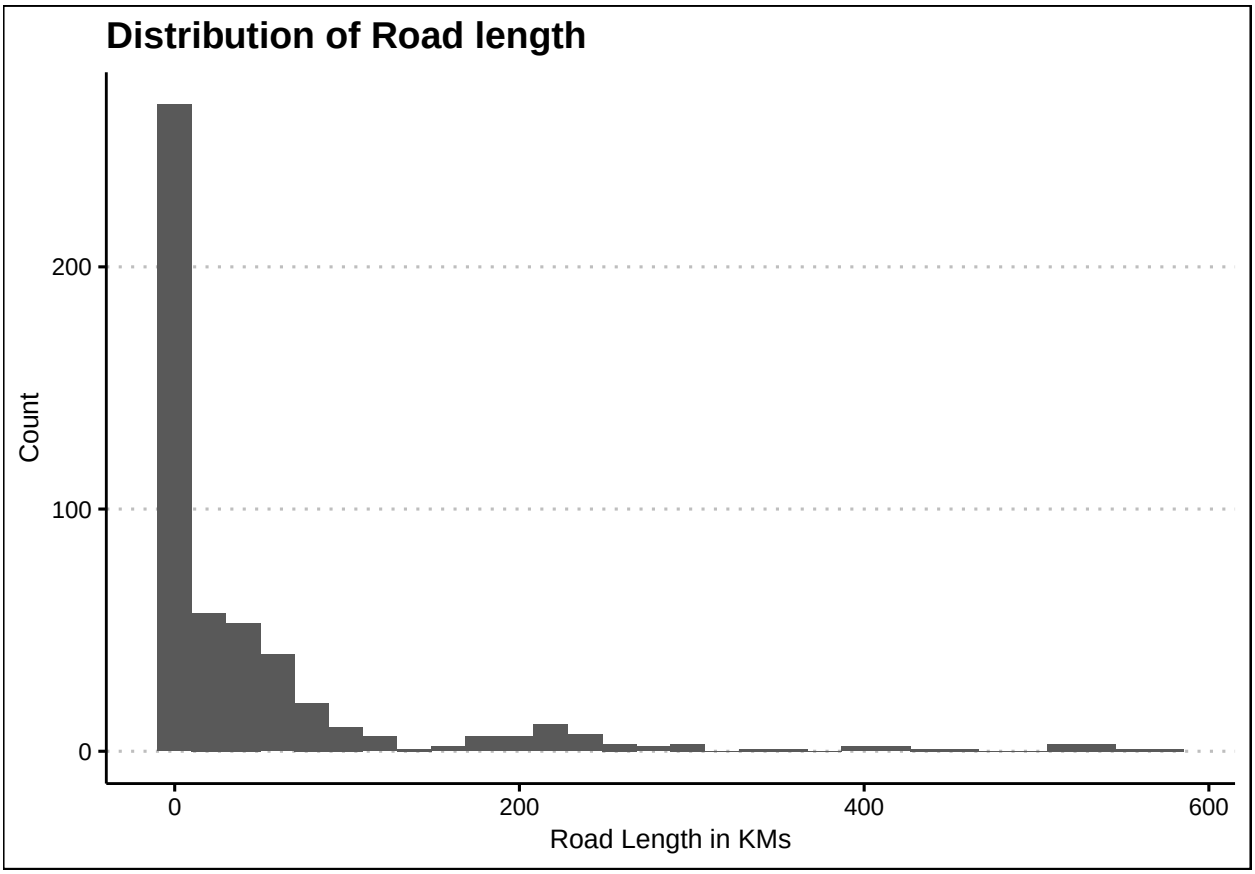



Figure 3: Distribution of Road length

I also slice the length of roads into chunks of 10km. The output shows that there are 270 roads that are at most 10 km further illustrating the skewness of the variable.

```
cut(roads$longueur_km, breaks = seq(from = min(roads$longueur_km),
                                    to = max(roads$longueur_km),by = 10)) %>%
  table()
```

```
## .
##    (0,10]   (10,20]   (20,30]   (30,40]   (40,50]   (50,60]   (60,70]   (70,80]
##       270        34        19        29        26        21        18        12
##   (80,90]  (90,100] (100,110] (110,120] (120,130] (130,140] (140,150] (150,160]
```

```
##         8         8         2         4         1         1         1         0
## (160,170] (170,180] (180,190] (190,200] (200,210] (210,220] (220,230] (230,240]
##         3         2         2         2         5         4         7         1
## (240,250] (250,260] (260,270] (270,280] (280,290] (290,300] (300,310] (310,320]
##         5         1         3         1         1         1         1         0
## (320,330] (330,340] (340,350] (350,360] (360,370] (370,380] (380,390] (390,400]
##         1         0         0         0         1         0         0         0
## (400,410] (410,420] (420,430] (430,440] (440,450] (450,460] (460,470] (470,480]
##         2         2         0         0         1         1         0         0
## (480,490] (490,500] (500,510] (510,520] (520,530] (530,540] (540,550] (550,560]
##         0         0         3         0         0         3         1         0
## (560,570]
##         0
```

## 5.2 What is the distribution of departmental roads?

Again, I use a histogram to visualise the distribution of departmental roads. Figure 4 shows that departmental roads are more evenly distributed with peaks between 25 kms to 50 kms. This distribution is closer to the normal distribution unlike the distribution of all roads that is heavily skewed to the right.

```
roads %>%
  filter(type_route == "departementale") %>%
  ggplot(mapping = aes(x = longueur_km)) +
  geom_histogram() +
  labs(x = "Road Length in KMs", y = "Count", title = "Distribution of Departmental Road length")
```

We can extend this analysis to all departments. Figure 5 shows that while national roads have a more even distribution, other road types are heavily skewed.

```
roads %>%
  ggplot(mapping = aes(x = longueur_km)) +
  geom_histogram() +
  labs(x = "Road Length in KMs", y = "Count", title = "Distribution of Road length by Road Type") +
  facet_wrap(~ factor(type_route), scales = "free")
```

## 5.3 To what extent does the number of vehicles using a road depend on the length of that road?

In this case, I test whether the relationship between the two variables is significant. I start by visualizing the two variables using a scatter plot because we have two quantitative variables. The scatter plot shows a general upward trend in the number of vehicles as the length of the roads increases (see Figure 6).

```
roads %>%
  ggplot(mapping = aes(x = n_vehicules_par_jour, y = longueur_km)) +
  geom_point(alpha = 0.5, size = 3) +
  geom_smooth(se = FALSE, method = "lm") + scale_color_tableau() +
  labs(title = "Number of Vehicles versus Road Length",
       x = "Number of Vehicles", y = "Road Length")
```

I also do a correlation test for the significance of the correlation between the length of the road (longueur_km) and the number of vehicles (n_vehicules_par_jour). This hypothesis is useful when dealing with two quantitative variables that have an approximate linear relationship as is the case here.

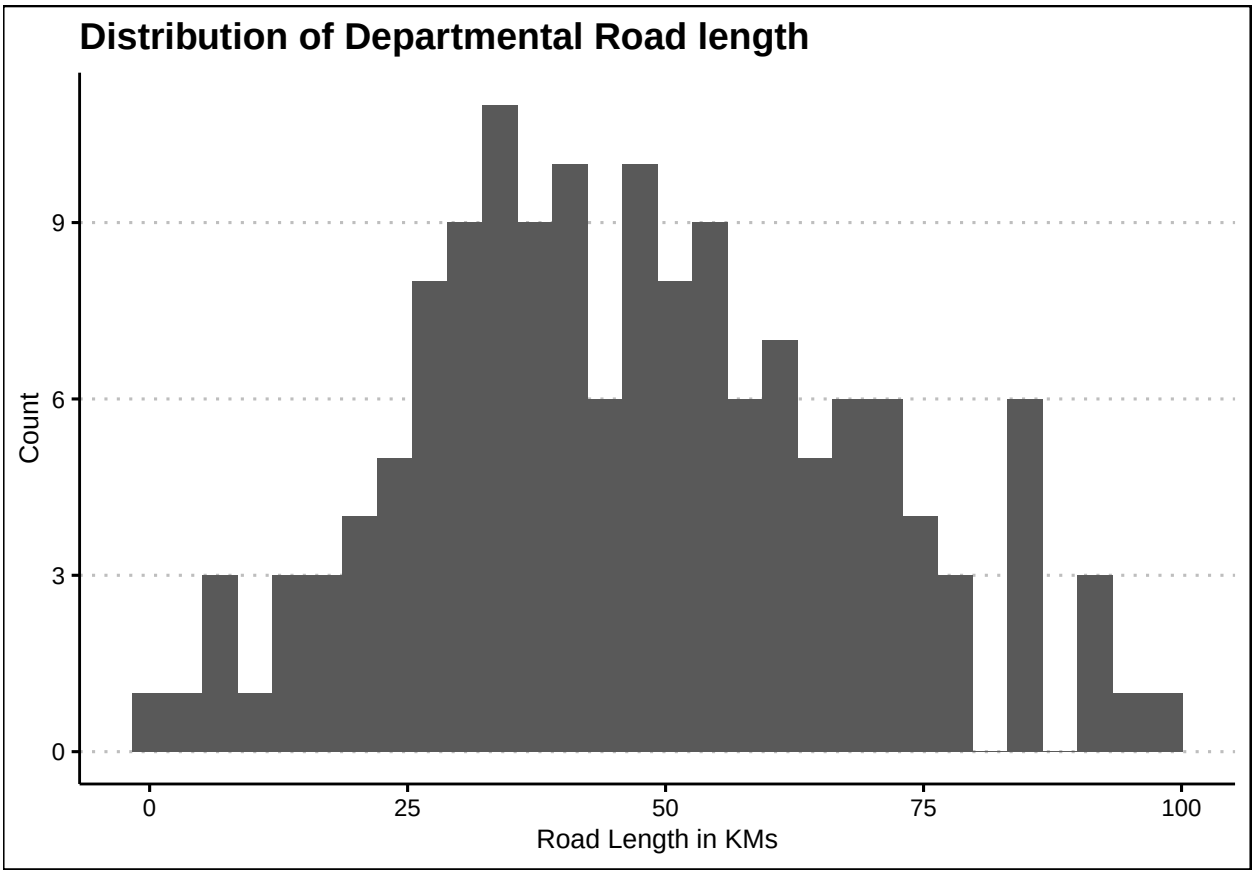



Figure 4: Distribution of Road length for Departmental Roads

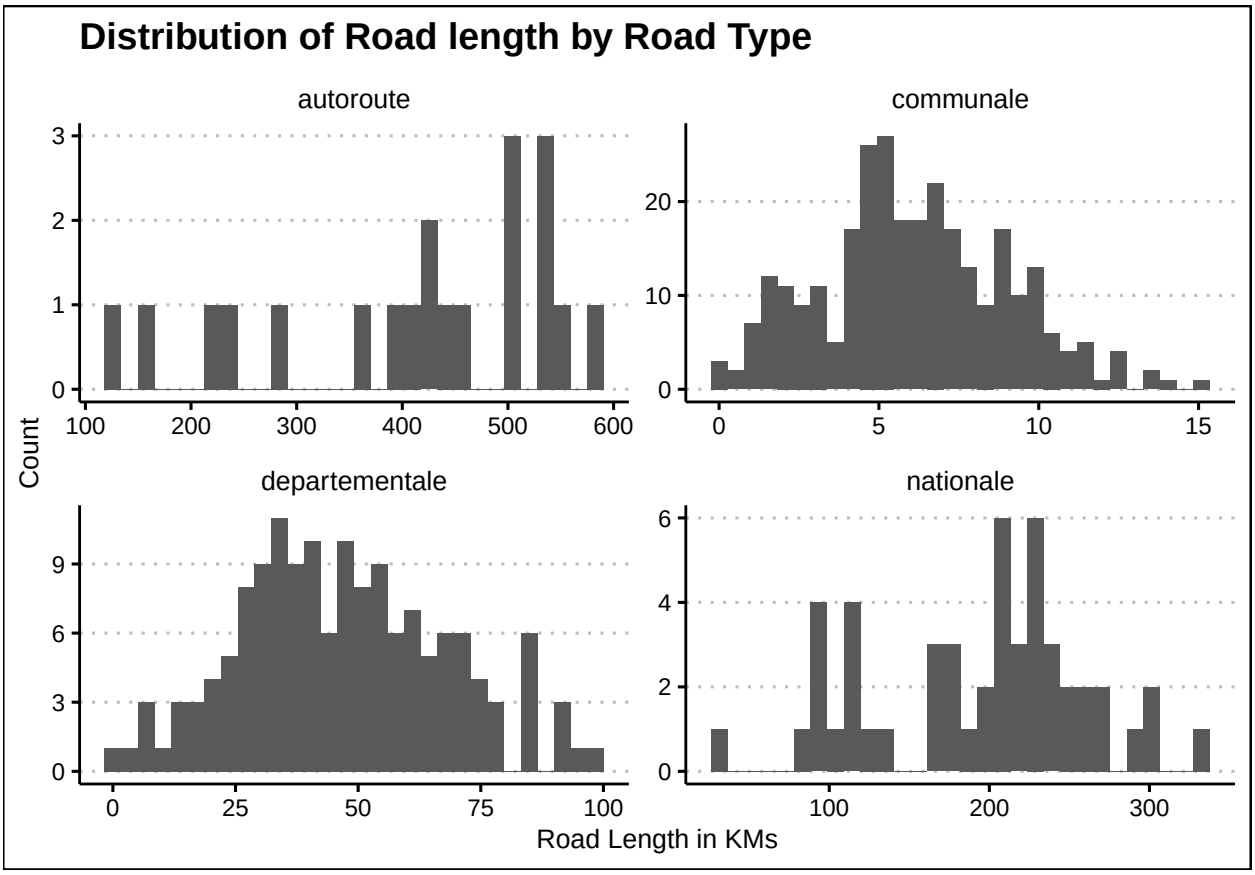



Figure 5: Distribution of Road length by Road Type

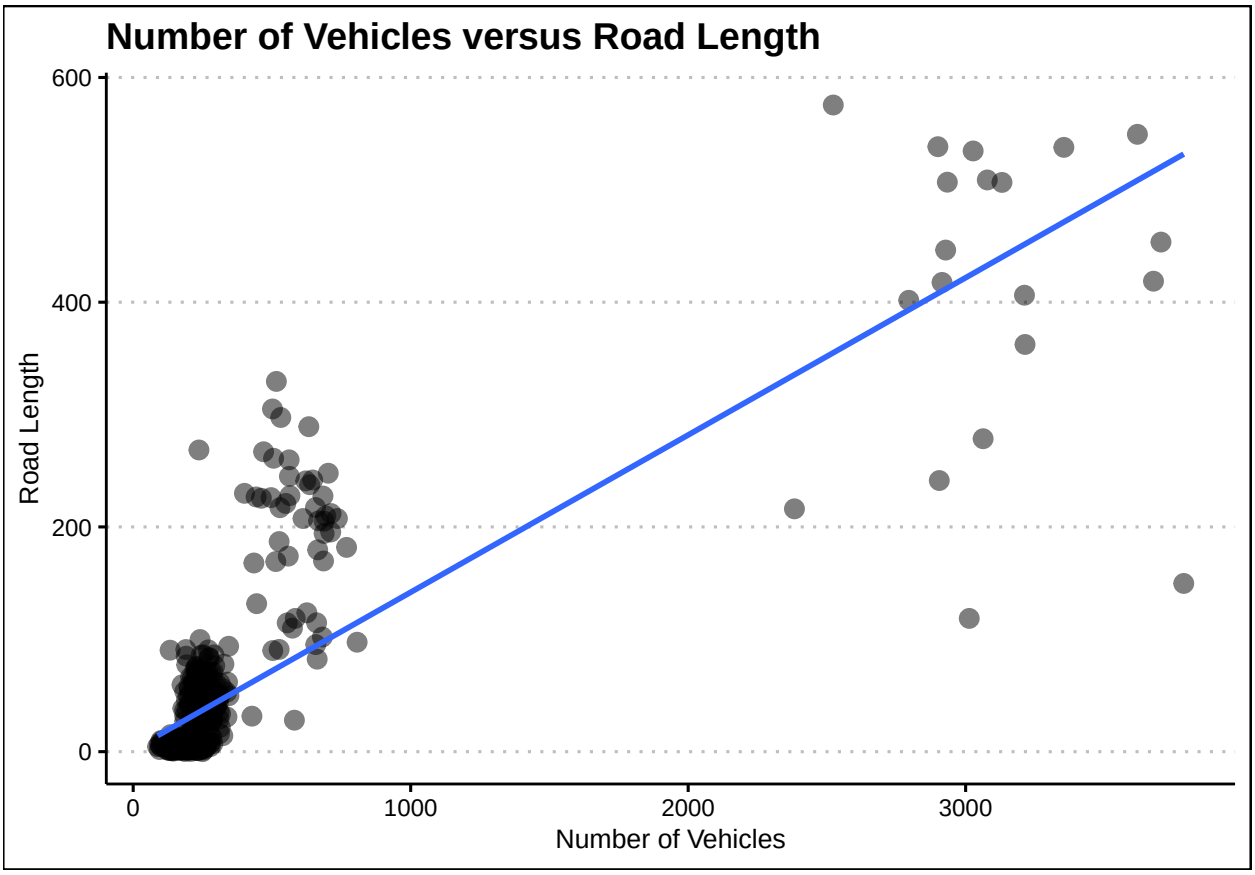



Figure 6: Number of Vehicles versus Road Length

```
with(roads, (cor.test(n_vehicules_par_jour, longueur_km)))
```

```
##
##  Pearson's product-moment correlation
##
## data:  n_vehicules_par_jour and longueur_km
## t = 33, df = 508, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.80 0.85
## sample estimates:
##  cor
## 0.83
```

The output shows that we reject the null hypothesis of no relationship between the road (longueur_km) and the number of vehicles (n_vehicules_par_jour) at 1% level of significance.

## 5.4 Is this ratio the same for the different road types?

In this section, I examine the correlation between the road (longueur_km) and the number of vehicles (n_vehicules_par_jour) for different road types. Figure () already shows that this correlation is different for different road types.

The extent of this relationship depends on the road type, with auto-route and departmental roads showing little correlation between road length and number of vehicles. Communal and national roads show a negative correlation between the length of the road and the number of vehicles. National roads have the highest negative correlation between road length and number of vehicles (see Figure 7).

```
roads %>%
  summarise(

    correlation = cor(n_vehicules_par_jour, longueur_km),

    .by = type_route
  ) %>%

  arrange(desc(correlation))
```

```
##       type_route correlation
## 1 departementale      0.0128
## 2      autoroute      0.0035
## 3      communale     -0.0284
## 4      nationale     -0.2360
```

```
roads %>%
  ggplot(mapping = aes(x = n_vehicules_par_jour, y = longueur_km, color = type_route)) +
  geom_point(alpha = 0.5, size = 3) + geom_smooth(se = FALSE, method = "lm") + scale_color_tableau() +
  labs(title = "Number of Vehicles versus Road Length and Road Type",
       x = "Number of Vehicles", y = "Road Length") + theme(legend.title = element_blank())
```

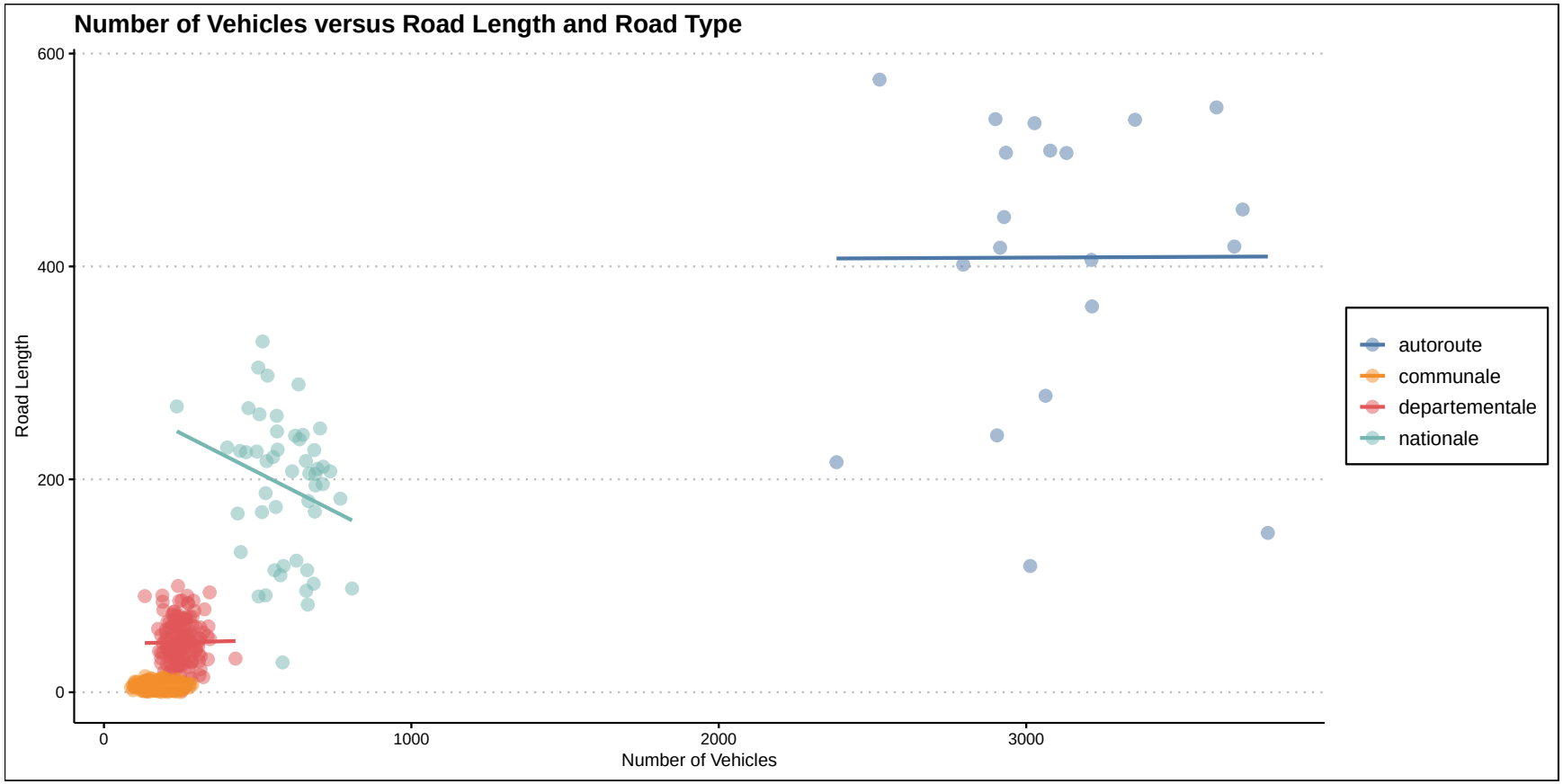



Figure 7: Number of Vehicles versus Road Length and Road Type

## 5.5 Are departmental roads shorter on average than national roads?

I create a summary table showing the mean and median road length for departmental and national roads. The analysis shows that departmental roads are,on average, about 5 times shorter than national roads.

```
roads %>%
  filter(type_route %in% c("departementale", "nationale")) %>%
  summarise(

    Mean_length = mean(longueur_km),
    Median_length = median(longueur_km),
    .by = type_route
  ) %>%

  set_names(names(.) %>% str_to_title()) %>%

  mutate(Type_route = str_to_title(Type_route)) %>%

  kbl(booktabs = TRUE, caption = "Road Lengths for National/Departmental Roads") %>%
  kable_classic(full_width = TRUE, latex_options = "hold_position")
```

Table 4: Road Lengths for National/Departmental Roads

| Type_route | Mean_length | Median_length |
|---|---|---|
| Nationale | 193 | 208 |
| Departementale | 47 | 46 |

I also carry out a t-test to examine whether the observed difference is statistically significant. We reject the null hypothesis at 1% significance level which means that there is a significant relationship between road type (national versus departmental) and the length of the road.

```
roads %>%
  filter(type_route %in% c("departementale", "nationale")) %>%
  mutate(type_route = factor(type_route, levels = c("departementale", "nationale"))) %>%
t.test(longueur_km ~ type_route, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  longueur_km by type_route
## t = -15, df = 52, p-value <2e-16
## alternative hypothesis: true difference in means between group departementale and group nationale is
## 95 percent confidence interval:
##  -166 -127
## sample estimates:
## mean in group departementale      mean in group nationale
##                           47                          193
```

## 5.6 Is there a difference in length depending on the number of lanes?

There is a notable difference in the length of roads depending on the number of lanes. Single lane roads are the rarest while 3 lane roads are the most abundant followed by 4 lane roads.

```
roads %>%
  mutate(n_voies = factor(n_voies)) %>%
```

```
  summarise(

    Mean_length = mean(longueur_km),
    Median_length = median(longueur_km),
    .by = n_voies
  ) %>%

  set_names(names(.) %>% str_to_title()) %>%

  mutate(N_voies = str_to_title(N_voies)) %>%

  kbl(booktabs = TRUE, caption = "Mean and Median Road Lengths by Number of lanes") %>%
  kable_classic(full_width = TRUE, latex_options = "hold_position")
```

Table 5: Mean and Median Road Lengths by Number of lanes

| N_voies | Mean_length | Median_length |
|---|---|---|
| 3 | 415 | 453.5 |
| 4 | 388 | 406.2 |
| 2 | 36 | 10.5 |
| 1 | 38 | 8.7 |

I visualise this difference in road lengths by the number of lanes in Figure 8. Again, the figure confirms that 3 lane and 4 lane roads dominate the data.

```
roads %>%
  ggplot(mapping = aes(y = longueur_km, x = factor(n_voies))) +
  geom_boxplot() + geom_jitter(shape = 'x') + scale_fill_tableau() +
  labs(title = "Road Length vs Number of Lanes", x = "Lanes", y = "Road length")
```

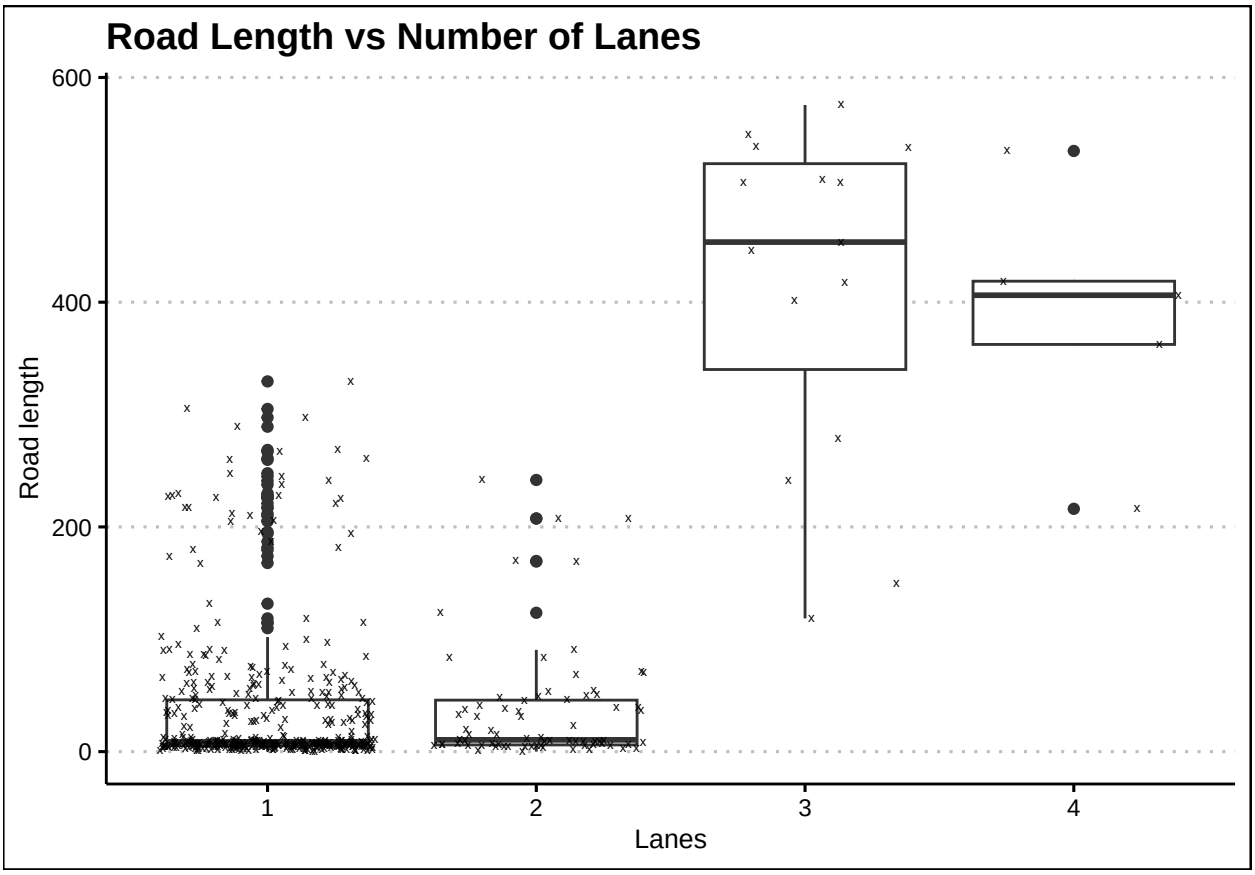
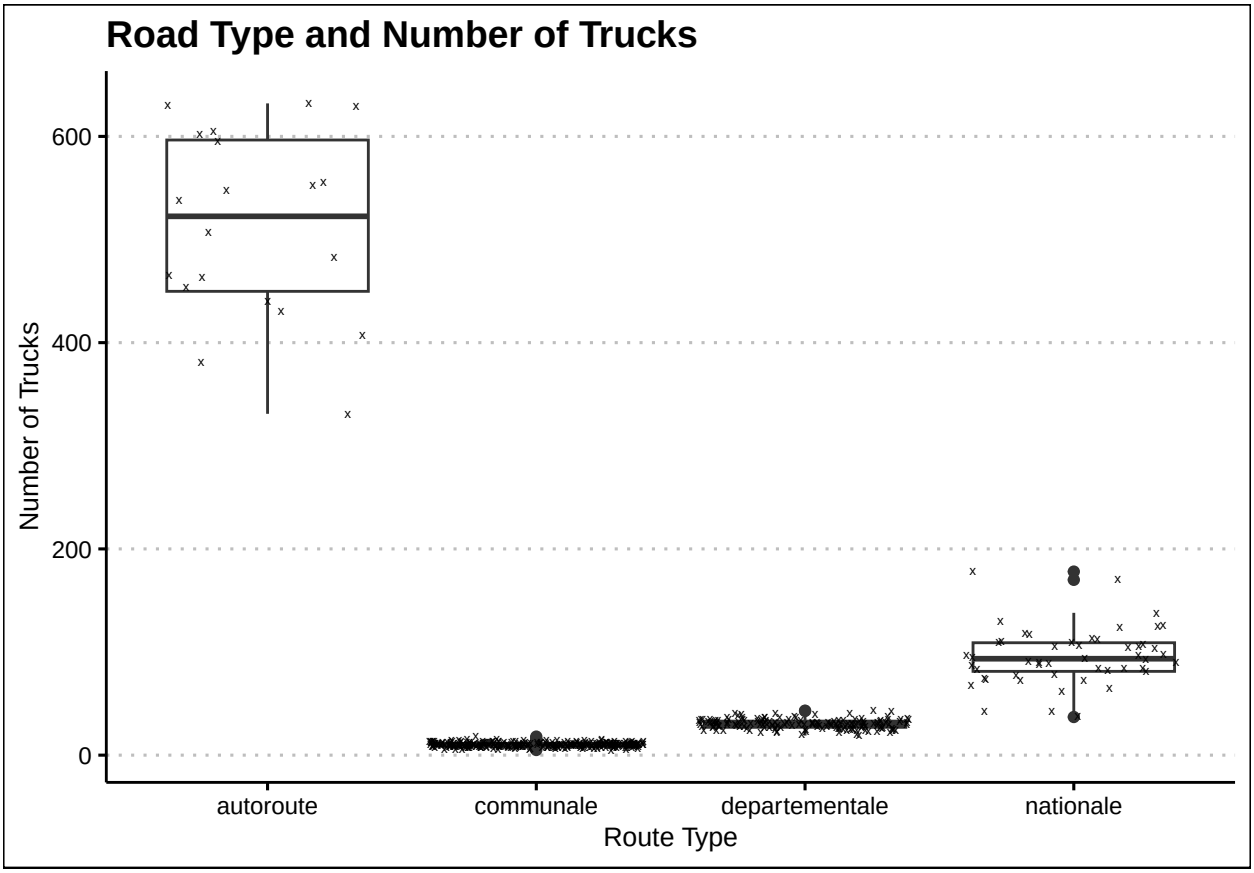



Figure 8: Road Length vs Number of Lanes

### 5.7 Which road types are dominated by trucks?

In this section, I examine the roads mostly used by trucks. I visualise the data in Figure () below. The graph shows that trucks are far more dominant on autoroutes followed at a distance by national roads.

```
roads %>%
  ggplot(mapping = aes(x = type_route, y = n_camions_par_jour)) +
  geom_boxplot() + geom_jitter(shape = "x") +
  labs(x = "Route Type", y = "Number of Trucks", title = "Road Type and Number of Trucks")
```




## 6 Conclusion

In this analysis, I utilised road usage data to explore several questions. The output shows the following;

1. The distribution of road length depends on the road type.
2. The usage of a road has a significant relationship with the road type.
3. Longer roads have, on average, more vehicles than shorter roads.
4. National roads are, on average, longer than other road types.
5. The usage of a road has a direct relationship with the number of lanes, with 3-lane and 4-lane roads the most used.

The analysis would be useful for traffic management.