CSE3020 - Data Visualisation

Lab Assignment 1

R Programming: TN District Data

SANJIT KUMAR
18BCE0715
DR. NALINI N
LAB - L13 + L14

Question

Consider the following transportation dataset of three districts. It describes district code, District name, Transport mode, Total population and people who drove alone.

D_code	District	Transport Mode	Pop_total	Selfdrive_total
45	Ranipet	Bicycle	73560	2414
45	Ranipet	Bike	1634923	42902
78	Thirupatthur		797818	21348
78	Thirupatthur		3865125	75246
78	Thirupatthur	Bicycle	42880	1088
78	Thirupatthur		7710301	399041
111	Vellore	Car	373402	13922
111	Vellore	Bicycle	27313	1075
111	Vellore	Bike	14525322	557036

Write R code for the below questions

- Create data frame for the above data
- How many observations of ‘district’ are missing from the data frame
- Count the number of self-drive in each district.
- Print max and min of pop_total.
- Derive new information/print "percentage of people who drove alone in all three districts" and also rank districts based on the % of people who used bicycle.

Source Code

```
# 18BCE0715 - Sanjit Kumar
# dataset creation


tn_transport_data <- data.frame(
  D_code = c(45,45,78,78,78,78,111,111,111),
  District = c(rep(c("Ranipet"),times=2),
rep(c("Thirupatthur"),times=4), rep(c("Vellore"),times=3)),
  Transport_Mode =
c("Bicycle","Bike",NA,NA,"Bicycle",NA,"Car","Bicycle","Bike"),
  Pop_total =
c(73560,1634923,797818,3865125,42880,7710301,373402,27313,14525322
),
  Selfdrive_total =
c(2414,42902,21348,75246,1088,399041,13922,1075,557036)
)


tn_transport_data


# total number of missing values
print("Total number of missing values:")
sum(is.na(tn_transport_data))


# max and min of population total
print("Maximum value of attribute Pop_total is:")
max(tn_transport_data$Pop_total)
print("Minimum value of attribute Pop_total is:")
min(tn_transport_data$Pop_total)


# percentage of people who drove alone in all three districts
print("percentage of people who drove alone in all three
districts")
```

```
percentage_of_selfdriving_people =
sum(tn_transport_data$Selfdrive_total)*100/
sum(tn_transport_data$Pop_total)

percentage_of_selfdriving_people


# districts ranked on % of bicycle users

print("districts ranked on % of bicycle users")

ranipet_subset = subset(tn_transport_data,District=="Ranipet")

thirupatthur_subset =
subset(tn_transport_data,District=="Thirupatthur")

vellore_subset = subset(tn_transport_data,District=="Vellore")


s1 <- tn_transport_data$Pop_total[tn_transport_data$Transport_Mode
== "Bicycle" & tn_transport_data$District == "Ranipet" ]*100/
sum(tn_transport_data$Pop_total[tn_transport_data$District ==
"Ranipet"])

s1

s2 <- tn_transport_data$Pop_total[tn_transport_data$Transport_Mode
== "Bicycle"  & !is.na(tn_transport_data$Transport_Mode) &
tn_transport_data$District == "Thirupatthur"]*100/
sum(tn_transport_data$Pop_total[tn_transport_data$District ==
"Thirupatthur"])

s2

s3 <- tn_transport_data$Pop_total[tn_transport_data$Transport_Mode
== "Bicycle" & tn_transport_data$District == "Vellore" ]*100/
sum(tn_transport_data$Pop_total[tn_transport_data$District ==
"Vellore"])

s3

df_bicycle_district_percent = data.frame(

  District=c("Ranipet","Thirupatthur","Vellore"),

  bicycle_percent = c(s1,s2,s3)

)

df_bicycle_district_percent

rank(df_bicycle_district_percent$bicycle_percent)
```

Output Screenshots

a. Data frame with given data

```
> source('~/Documents/VIT_DOC/vit_semester_6/B2 - Data Visualisation/lab/submission1/18BCE0715_Assignment1.R')
> tn_transport_data <- data.frame(
+   D_code = c(45,45,78,78,78,78,111,111,111),
+   District = c(rep(c("Ranipet"),times=2), rep(c("Thirupatthur"),times=4), rep(c("Vellore"),times=3)),
+   Transport_Mode = c("Bicycle","Bike",NA,NA,"Bicycle",NA,"Car","Bicycle","Bike"),
+   Pop_total = c(73560,1634923,797818,3865125,42880,7710301,373402,27313,14525322),
+   Selfdrive_total = c(2414,42902,21348,75246,1088,399041,13922,1075,557036)
+ )
> tn_transport_data
  D_code     District Transport_Mode Pop_total Selfdrive_total
1     45      Ranipet        Bicycle     73560            2414
2     45      Ranipet           Bike   1634923           42902
3     78 Thirupatthur           <NA>    797818           21348
4     78 Thirupatthur           <NA>   3865125           75246
5     78 Thirupatthur        Bicycle     42880            1088
6     78 Thirupatthur           <NA>   7710301          399041
7    111      Vellore            Car    373402           13922
8    111      Vellore        Bicycle     27313            1075
9    111      Vellore           Bike  14525322          557036
```

b. Missing data from frame

```
[1] "Total number of missing values:"
> sum(is.na(tn_transport_data))
[1] 3
```

c. Max and Min of Pop_Total

```
> # max and min of population total
> print("Maximum value of attribute Pop_total is:")
[1] "Maximum value of attribute Pop_total is:"
> max(tn_transport_data$Pop_total)
[1] 14525322
> print("Minimum value of attribute Pop_total is:")
[1] "Minimum value of attribute Pop_total is:"
> min(tn_transport_data$Pop_total)
[1] 27313
```

d.
- Percentage of people who drive bicycle in all three districts

```
[1] "percentage of people who drove alone in all three districts"
> percentage_of_selfdriving_people = sum(tn_transport_data$Selfdrive_total)*100/sum(tn_transport_data$Pop_total)
> percentage_of_selfdriving_people
[1] 3.83493
```

- Districts ranked on % of bicycle users

```
[1] "districts ranked on % of bicycle users"
> ranipet_subset = subset(tn_transport_data,District=="Ranipet")
> thirupatthur_subset = subset(tn_transport_data,District=="Thirupatthur")
> vellore_subset = subset(tn_transport_data,District=="Vellore")
> s1 <- tn_transport_data$Pop_total[tn_transport_data$Transport_Mode == "Bicycle" & tn_transport_data$District == "R
0/sum(tn_transport_data$Pop_total[tn_transport_data$District == "Ranipet"])
> s1
[1] 4.305574
> s2 <- tn_transport_data$Pop_total[tn_transport_data$Transport_Mode == "Bicycle"  & !is.na(tn_transport_data$Transp
n_transport_data$District == "Thirupatthur"]*100/sum(tn_transport_data$Pop_total[tn_transport_data$District == "Thir
> s2
[1] 0.3453574
> s3 <- tn_transport_data$Pop_total[tn_transport_data$Transport_Mode == "Bicycle" & tn_transport_data$District == "V
0/sum(tn_transport_data$Pop_total[tn_transport_data$District == "Vellore"])
> s3
[1] 0.182989
> df_bicycle_district_percent = data.frame(
+   District=c("Ranipet","Thirupatthur","Vellore"),
+   bicycle_percent = c(s1,s2,s3)
+ )
> df_bicycle_district_percent
      District bicycle_percent
1      Ranipet       4.3055740
2 Thirupatthur       0.3453574
3      Vellore       0.1829890
> rank(df_bicycle_district_percent$bicycle_percent)
[1] 3 2 1
>
```