Міністерство освіти та науки України

Національний університет «Львівська політехніка»

Інститут комп’ютерних наук та інформаційних технологій

Кафедра інформаційних систем та мереж

Звіт

про виконання лабораторної роботи № 4

з дисципліни Візуалізація даних

на тему «Візуалізація категоріальних та числових даних засобами RStudio»

Варіант 8

Виконала студентка групи СА-41

Дугіна С. О.

Прийняв к.т.н., доцент

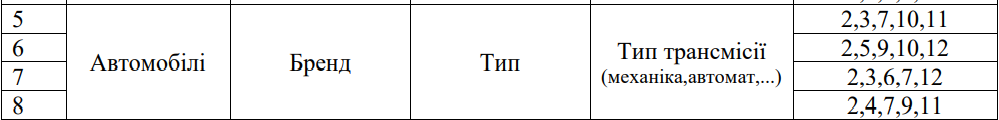
Андруник В. А.

Львів 2022

**Мета роботи:**  набуття практичних навичок опрацювання та візуалізації категоріальних та числових даних в середовищі RStudio.

**Хід виконання**

Table

Description automatically generated

Обраний датасет: [Ford Car Price Prediction](https://www.kaggle.com/datasets/adhurimquku/ford-car-price-prediction?resource=download). Це датасет із автомобілями Ford та цінами на них. У датасеті є наступні елементи: модель (model), рік (year), ціна (price), вид передач (transmission), пробіг (mileage), тип палива (fuelType), податок (tax), mgp (miles per gallon) та розмір двигуна (engineSize). Усі ці фактори можуть впливати на ціну. Датасет має наступний вигляд:

Table

Description automatically generated

*Рисунок 1. Вигляд датасету в RStudio*

Види передач, що винесені в датасет: Automatic (автоматична коробка передач), Manual (механічна коробка передач), Semi-Auto (напівавтоматична коробка передач). Типи палива: Petrol (бензин), Diesel (дизель), Hybrid (гібрид), Electric (електричний), Other (інший).

Необхідно створити наступні діаграми: Lollipop, Donut, Treemap, Parallel plot та Scatter plot.

Першою створимо Lollipop. Дана діаграма працює як barplot, проте подається в іншому вигляді (стовпець змінюється на пряму та крапку). Оскільки ми працюємо із великою кількістю даних, то ця діаграма є легшою для сприйняття. Наприклад, таким чином виглядатиме барплот:

df %>%

ggplot(aes(x = price, y = transmission)) +

geom\_bar(aes(fill = transmission), position = "dodge", size=0.5, stat="identity", color="black") +

scale\_fill\_brewer(palette = 'PuBu') +

theme\_bw()

Chart, bar chart

Description automatically generated

*Рисунок 2. Barplot для заданого датасету*

Lollipop виглядатиме наступним чином:

df %>%

mutate(transmission = paste(transmission, fuelType)) %>%

group\_by(transmission) %>%

summarise(price=round(mean(price))) %>%

arrange(price) %>%

mutate(order = seq\_len(length(transmission))) %>%

ggplot(aes(x = reorder(transmission, order), y = price)) +

geom\_segment(aes(xend=transmission, y=0, yend=price),color="darkslateblue",lwd=0.7) +

geom\_point(color="darkslateblue", size=5, shape=19) +

geom\_label(aes(x = transmission, y = price, label=price), vjust = -0.4) +

theme\_bw() +

coord\_flip()

Chart, scatter chart

Description automatically generated

*Рисунок 3. Lollipop діаграма для заданого датасету*

По-перше, бачимо, що окрім типу передач ми додаємо ще тип палива. І, відповідно, таким чином можна бачити, що, наприклад, найнижча ціна є в автомобілів з механічною коробкою передач та на бензині. Найдорожчими є напівавтомати на гібриді. Варто також зауважити, що в якості «цін» було взято середні значення цін усіх моделей з відповідними характеристиками (тип палива та коробки передач).

Наступна діаграма – Donut. За допомогою цієї діаграми, а також Pie Chart, можна візуалізувати пропорції даних. Для прикладу розглянемо типи коробок передач та ціни на автомобілі з конкретними типами:

df %>%

group\_by(transmission) %>%

summarise(price = round(mean(price))) %>%

mutate(size=4.5) %>%

ggplot(aes(x = size, y = price, fill=transmission)) +

geom\_col(color="white", lwd=1) +

geom\_label(aes(label = price),position= position\_stack(vjust = 0.5)) +

scale\_fill\_brewer(palette = 'PuBu') +

coord\_polar(theta = "y") + xlim(c(3, 5)) +

theme\_void()

Chart, sunburst chart

Description automatically generated

*Рисунок 4. Donut діаграма для заданого датасету*

Знову взято середні значення цін. Бачимо, що найдешевшими є автомобілі із механічною коробкою передач, що відповідає і попередній діаграмі. Проте також бачимо, що дорожчими є саме автомобілі із автоматичними коробками передач, а не напівавтоматичними.

Також було розроблено Pie Chart для наглядності:

df %>%

group\_by(transmission) %>%

summarise(price = round(mean(price))) %>%

ggplot(aes(x = "", y = price, fill=transmission)) +

geom\_bar(stat="identity", color="white", lwd=1) +

geom\_label(aes(label = price), position= position\_stack(vjust = 0.5)) +

scale\_fill\_brewer(palette = 'PuBu') +

coord\_polar(theta = "y") +

theme\_void()

Chart, pie chart

Description automatically generated

*Рисунок 5. Pie Chart діаграма для заданого датасету*

Як бачимо, результати ті самі. Різниця лише у вигляді.

Наступною діаграмою розробимо Treemap:

# treemap

library(treemap)

treemap(df,

index="model",

vSize="price",

vColor="engineSize",

palette = "PuBu",

type="index")

Chart, treemap chart

Description automatically generated

*Рисунок 6. Treemap діаграма для заданого датасету*

Treemap – це узагальнення теплової карти (heatmap). В даному випадку розмір кожного прямокутника вказує на ціну певної моделі. Бачимо, що найдорожчими моделями є Fiesta, Focus та Kuga.

Розробимо Parallel Plot. За допомогою даної діаграми можна візуалізувати багатовимірні дані. Застосуємо бібліотеку GGally. Розглянемо ціни та моделі автомобілів, а також певні фактори, наприклад рік, ціну, пробіг та податки. Також можна розбити на окремі діаграми для кожної характеристики, але для загальної картини розглянемо одну діаграму:

df %>%

mutate(price = mean(price)) %>%

slice\_sample(n=100) %>%

ggparcoord(

groupColumn = "model", columns=c(2,7,9),

splineFactor = 10, showPoints = T,

) + scale\_color\_brewer(palette = "Paired") + theme\_bw()

Chart, line chart

Description automatically generated

*Рисунок 7. Parallel plot для заданого датасету*

Останньою діаграмою розглянемо Scatter plot:

scatterplot <- ggplot(data=df) + geom\_point(aes(x=year, y=model,

size=price,color = transmission), alpha=0.7) + xlim(NA,2025)

scatterplot + ggtitle("something")

Chart, scatter chart

Description automatically generated

*Рисунок 8. Scatter plot для заданого датасету*

Для кращого візуального сприйняття було обрано скоротити роки до 2025.Таким чином, розмір точок вказує на ціну, колір – на тип коробки передач. Бачимо для кожної моделі в залежності від років картину цін.

**Висновок**

У даній лабораторній роботі було розроблено ряд діаграм для датасету Ford Car Price Prediction, а саме: Lollipop, Donut, Treemap, Parallel Plot, Scatter Plot. Окрім того, для порівняння було створено Barplot та Pie Chart.