KRAFTWER KORP SALES ANALYSIS

Problem Statement

Kraftwerk Korp is Atlas AI's new customer. They have 500 stores spread across Nairobi selling 3 different products: Bits, Bobs and Widgets. Kraftwerk has shared with you as a data analyst their sales data which include the location of each store and total sales volume for each product and expect you as a data analyst to help them answer two key questions:

1.What is the relationship between store performance and location?
2.Based on that relationship, where should Kraftwerk build more stores?

Data Analysis Guide

We will use the Google Data Analytics framework to guide us through the analysis:

1.**Ask** - **Business Challenge/Objective/Question**

The key business questions that Kraftwerk wants to answer include:

i)What are the top 5 neighborhoods in terms of revenue?

ii)What are the top 5 stores in terms of sales quantity?

iii)What is the best selling item in terms of both sales quanity and revenue?

iv)Which part of Nairobi has the highest concentration of stores?

2.**Prepare** - **Data generation, collection, storage and management**

We will load the sales data into R studio session and explore it to understand its structure and properties.

3.**Process** - **Data cleaning and transformation**

The data cleaning in this case will involve:

-inspecting for NAs/missing values

-formatting column names

-checking for duplicate records

Data transformation will entail:

-calculating revenue per product

-sorting sales and revenue data

-calculating total revenue and sales for all the products

4.**Analyze** - **Data exploration, visualization and analysis**

The analysis phase will involve creating static visualizations such as bar plots, box plots and drawing conclusions.

5.**Share** - **Communicating and interpreting results**

We will use R shiny tool to help bring the data to life and answer the key business questions.

6.**Act** - **Putting insights to work**

Finally, we will make data-driven recommendations to Kraftwerk on store performance and expansion strategies

Analysis references:

The sales analysis github link: Krafter Korp sales analysis

The sales dashboard github link: Krafter Korp sales dashboard

Part 1: General EDA - Getting to Know the Data

1.1 Importing Required Packages

If the required libraries are not installed in your R studio session, the following script will install and load the libraries

```r
#script installs required packages if not installed in the R session
if(!require(rmarkdown)) install.packages("rmarkdown")
if(!require(tinytex)) install.packages("tinytex")
if(!require(tidyverse)) install.packages("tidyverse")
if(!require(skimr)) install.packages("skimr")
if(!require(readr)) install.packages("readr")
if(!require(extrafont)) install.packages("extrafont")
if(!require(extrafont)) install.packages("ggthemes")
if(!require(ggtext)) install.packages("ggtext")
if(!require(ggtext)) install.packages("leaflet")


#load the packages
library(rmarkdown)
library(tinytex)
library(tidyverse)
library(skimr)
library(readr)
library(extrafont)
library(ggthemes)
library(ggtext)
library(leaflet)

#font_import()
#loadfonts(device = "win")
```

1.2 Data Preparation

1.2.1 Load the data

```r
data_path = "./www/Kraftwerk_korp_sales.csv"

sales_data = read_csv(data_path)

#script to have the data overview
head(sales_data)
```

id	neighborhood	shop_name	y	x	bits_qty	bobs_bqty	widgets_qty
<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Mlolongo	MG 4 phase 3 dreamland	36.95100	-1.392640	6738	3256	2578
2	Umoja 2 and 3	Green Grocer Shop	36.90144	-1.282295	6369	3679	750
3	Shauri Moyo	RAMANI SHOP 7068	36.84931	-1.296621	7681	2052	7361
4	Buruburu	DAUDI'S ENTERPRISES	36.87552	-1.276471	4473	2961	3466
5	Kangemi	Seraben supermarket-EA	36.74876	-1.266479	2629	9458	7218
6	Kayole	Eagle Services	36.91203	-1.276038	9706	3896	3865

6 rows

1.2.2 Inspect the data

The sales data has 501 rows, 8 columns and no missing values.

```r
#creating function to understand the data structure
data_structure <- function(df){

  #column data types
  str(df)

  #summary statistics mean,median,max,min
  summary(df)

  #checks for missing values,sd,p0,p25,p50,p75,p100
  skim(df)
}

#applying the function to the data
data_structure(sales_data)
```

```
## spec_tbl_df [501 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id          : num [1:501] 1 2 3 4 5 6 7 8 9 10 ...
##  $ neighborhood: chr [1:501] "Mlolongo" "Umoja 2 and 3" "Shauri Moyo" "Buruburu" ...
##  $ shop_name   : chr [1:501] "MG 4 phase 3 dreamland" "Green Grocer Shop" "RAMANI SHOP 7068" "DAUDI'S ENTERPRI
SES" ...
##  $ y           : num [1:501] 37 36.9 36.8 36.9 36.7 ...
##  $ x           : num [1:501] -1.39 -1.28 -1.3 -1.28 -1.27 ...
##  $ bits_qty    : num [1:501] 6738 6369 7681 4473 2629 ...
##  $ bobs_bqty   : num [1:501] 3256 3679 2052 2961 9458 ...
##  $ widgets_qty : num [1:501] 2578 750 7361 3466 7218 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   neighborhood = col_character(),
##   ..   shop_name = col_character(),
##   ..   y = col_double(),
##   ..   x = col_double(),
##   ..   bits_qty = col_double(),
##   ..   bobs_bqty = col_double(),
##   ..   widgets_qty = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data summary

Name	df
Number of rows	501
Number of columns	8
_______________________	
Column type frequency:	
character	2
numeric	6
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
neighborhood	0	1	3	29	0	81	0
shop_name	0	1	4	50	0	501	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	251.00	144.77	1.00	126.00	251.00	376.00	501.00	▇▇▇▇▇
y	0	1	36.88	0.09	36.63	36.84	36.89	36.92	37.13	▁▂▇▂▁
x	0	1	-1.26	0.08	-1.51	-1.30	-1.27	-1.23	-1.04	▁▁▇▂▁
bits_qty	0	1	4878.81	2938.62	5.00	2302.00	4819.00	7226.00	9995.00	▇▇▇▇▇
bobs_bqty	0	1	5175.23	2872.37	4.00	2810.00	5117.00	7710.00	9964.00	▆▇▆▆▇
widgets_qty	0	1	4973.45	2905.55	12.00	2351.00	4930.00	7598.00	9994.00	▇▇▇▇▇

1.2.1 Checking for duplicates

There are no duplicate data entries in the dataset

```r
sum(duplicated(sales_data))
```

```
## [1] 0
```

1.3 Data Cleaning and Transformation

Renaming the columns

```r
#renaming columns
sales_data <- sales_data %>%
  rename(
    'shop_location' = neighborhood,
    'lon' = y,
    'lat' = x
  )
head(sales_data)
```

id	shop_location	shop_name	lon	lat	bits_qty	bobs_bqty	widgets_qty
<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Mlolongo	MG 4 phase 3 dreamland	36.95100	-1.392640	6738	3256	2578
2	Umoja 2 and 3	Green Grocer Shop	36.90144	-1.282295	6369	3679	750
3	Shauri Moyo	RAMANI SHOP 7068	36.84931	-1.296621	7681	2052	7361
4	Buruburu	DAUDI'S ENTERPRISES	36.87552	-1.276471	4473	2961	3466
5	Kangemi	Seraben supermarket-EA	36.74876	-1.266479	2629	9458	7218
6	Kayole	Eagle Services	36.91203	-1.276038	9706	3896	3865

6 rows

Calculating the products' revenue for each store

```r
sales_revenue_data <- sales_data %>%
  mutate(
    bits_revenue = bits_qty*100,
    bobs_revenue = bobs_bqty*150,
    widgets_revenue = widgets_qty*300,
    total_revenue = bits_revenue+bobs_revenue+widgets_revenue,
    total_sales =  bits_qty+bobs_bqty+widgets_qty

  )
head(sales_revenue_data)
```

id	shop_location	shop_name	lon	lat	bits_qty	bobs_bqty	widgets_qty
<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Mlolongo	MG 4 phase 3 dreamland	36.95100	-1.392640	6738	3256	2578
2	Umoja 2 and 3	Green Grocer Shop	36.90144	-1.282295	6369	3679	750
3	Shauri Moyo	RAMANI SHOP 7068	36.84931	-1.296621	7681	2052	7361
4	Buruburu	DAUDI'S ENTERPRISES	36.87552	-1.276471	4473	2961	3466
5	Kangemi	Seraben supermarket-EA	36.74876	-1.266479	2629	9458	7218
6	Kayole	Eagle Services	36.91203	-1.276038	9706	3896	3865

6 rows | 1-8 of 13 columns

Saving the processed and cleaned sales data

```r
write.csv(sales_revenue_data,"./www/processed_sales_data.csv",row.names = FALSE)
```

converting sales data to long format for use in dashboard

```r
#creating long data format for sales quantity for use in dashboard
salesqty_dashboard_data <- sales_revenue_data %>%
  select(id:widgets_qty)%>%
  gather(product, sales_qty,bits_qty:widgets_qty) %>%
  mutate(product=case_when(
    product == "bits_qty" ~ "bits",
    product == "bobs_bqty" ~ "bobs",
    product == "widgets_qty" ~ "widgets"
  ))
head(salesqty_dashboard_data)
```

id	shop_location	shop_name	lon	lat	product	sales_qty
<dbl>	<chr>	<chr>	<dbl>	<dbl>	<chr>	<dbl>
1	Mlolongo	MG 4 phase 3 dreamland	36.95100	-1.392640	bits	6738
2	Umoja 2 and 3	Green Grocer Shop	36.90144	-1.282295	bits	6369
3	Shauri Moyo	RAMANI SHOP 7068	36.84931	-1.296621	bits	7681
4	Buruburu	DAUDI'S ENTERPRISES	36.87552	-1.276471	bits	4473
5	Kangemi	Seraben supermarket-EA	36.74876	-1.266479	bits	2629
6	Kayole	Eagle Services	36.91203	-1.276038	bits	9706

6 rows

```r
#creating long data format for sales revenue for use in dashboard
salesrevenue_dashboard_data <- sales_revenue_data %>%
  select(bits_revenue:widgets_revenue)%>%
  gather(product, sales_revenue,bits_revenue:widgets_revenue) %>%
  mutate(product=case_when(
    product == "bits_revenue" ~ "bits",
    product == "bobs_revenue" ~ "bobs",
    product == "widgets_revenue" ~ "widgets"
  )) %>% select(sales_revenue)
head(salesrevenue_dashboard_data)
```

sales_revenue
<dbl>
673800
636900
768100
447300
262900
970600

6 rows

```r
#merging the datasets by column
final_dashboard_data <- cbind(salesqty_dashboard_data,salesrevenue_dashboard_data)
head(final_dashboard_data)
```

	id	shop_location	shop_name	lon	lat	product	sales_qty	sales_revenue
	<dbl>	<chr>	<chr>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>
1	1	Mlolongo	MG 4 phase 3 dreamland	36.95100	-1.392640	bits	6738	673800
2	2	Umoja 2 and 3	Green Grocer Shop	36.90144	-1.282295	bits	6369	636900
3	3	Shauri Moyo	RAMANI SHOP 7068	36.84931	-1.296621	bits	7681	768100
4	4	Buruburu	DAUDI'S ENTERPRISES	36.87552	-1.276471	bits	4473	447300
5	5	Kangemi	Seraben supermarket-EA	36.74876	-1.266479	bits	2629	262900
6	6	Kayole	Eagle Services	36.91203	-1.276038	bits	9706	970600

6 rows

```r
#merging the datasets by column
write.csv(final_dashboard_data,"./www/dashboard_data.csv",row.names = FALSE)
```

calculating cumulative sales by neighborhood

```r
sales_revenue_neighborhood <- sales_revenue_data %>%
  group_by(shop_location) %>%
  summarise(
    neighborhood_revenue = sum(total_revenue)
  )
head(sales_revenue_neighborhood)
```

shop_location	neighborhood_revenue
<chr>	<dbl>
Athi River	20498750
Bahati	1077650
Buruburu	20370850
CBD	2779500
Chokaa	4526700
Dagoretti Corner	8101350

6 rows

Part 2: General EDA - Answering Business Questions

We will set a plotting theme for our plots

```r
#setting theme
f1<-"Comic Sans MS"

custom_theme<-function()
{
  #setting font
Font<-"Comic Sans MS"

theme_economist() %+replace%

  theme(
    ########grid elements########
    panel.grid.major = element_blank(),    #strip major gridlines
    panel.grid.minor = element_blank(),    #strip minor gridlines
     axis.ticks = element_blank(),          #strip axis ticks
    axis.line = element_blank(),
    ######text elements###########
    plot.title = element_markdown(             #title
                   family = Font,            #set font family
                   size = 13,                #set font size
                   face = 'bold',            #bold typeface
                   hjust = 0,                #left align
                   vjust = 2),               #raise slightly

      plot.subtitle = element_markdown(          #subtitle
                   family = Font,            #font family
                   size = 12,hjust = 0),               #font size

      plot.caption = element_text(           #caption
                   family = Font,            #font family
                   size = 8.7,
                   hjust = 0
                   ),               #right align

      axis.title = element_text( family = Font,size = 11),               #font size

      axis.text = element_text( family = Font,size = 11),               #font size

      legend.text = element_text( family = Font,size = 11),
      legend.title = element_text( family = Font,size = 11),
      legend.position = "right",
      plot.background = element_rect(fill = "white", color = NA)


    ##########plot################

  )
}
```

Q1 - What are the top 5 neighborhoods in terms of revenue?

Insights:

Thika leads with the highest revenue generation at Kes 78,133,900 followed by Kayole (Kes 76,688,100), Kibera (Kes 50,975,000), Huruma (Kes 39,947,950) and Ruiru (Kes 39,075,400).

```r
#filtering top 5 neighborhoods in terms of revenue
 top5_neighborhoods <-  sales_revenue_neighborhood %>%
  arrange(desc(neighborhood_revenue)) %>%
  top_n(5)
```

```
## Selecting by neighborhood_revenue
```

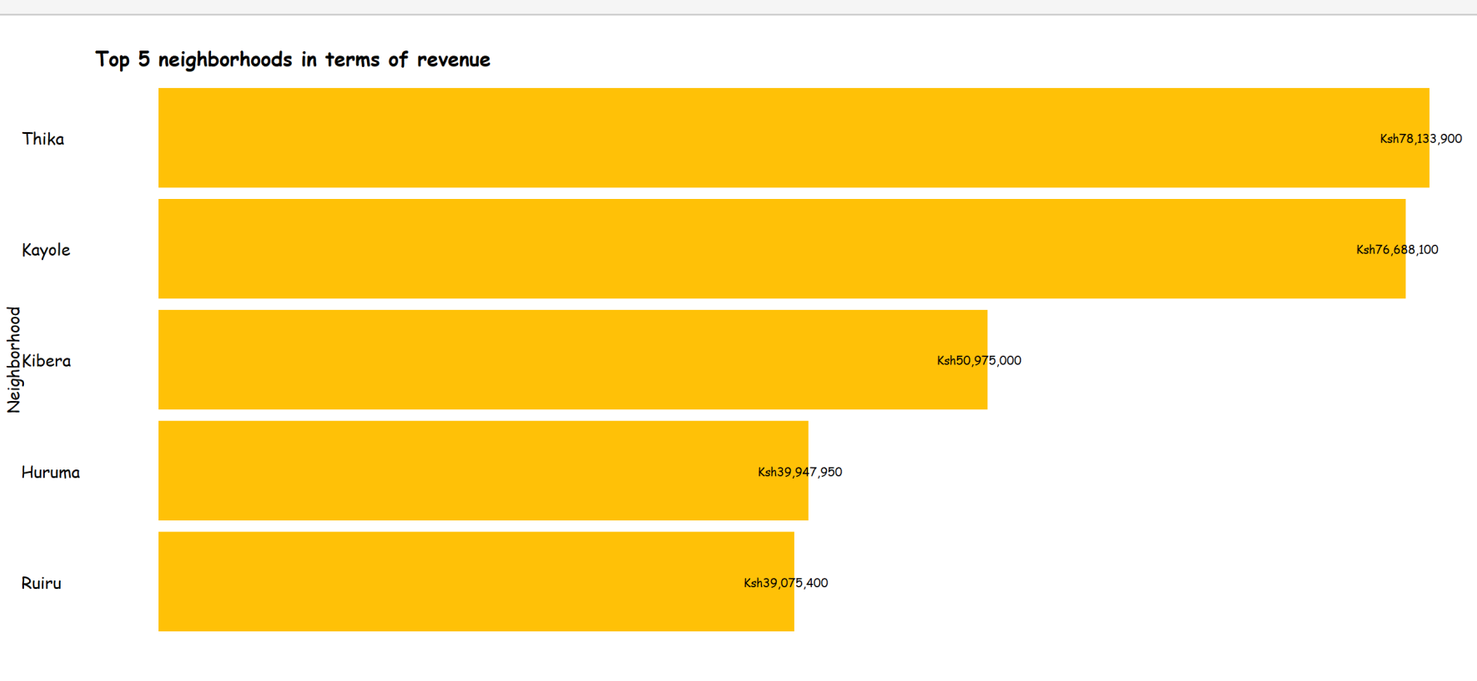
```r
#saving dataset for dashboard
write.csv(top5_neighborhoods,"./www/top5_neighborhoods_revenue.csv",row.names = FALSE)

#inspecting data
head(top5_neighborhoods)
```

shop_location	neighborhood_revenue
<chr>	<dbl>
Thika	78133900
Kayole	76688100
Kibera	50975000
Huruma	39947950
Ruiru	39075400

5 rows

Plotting top 5 neighborhoods in terms of revenue

```r
top5_neighborhoods %>%
  ggplot(aes(x=reorder(shop_location,neighborhood_revenue),y=neighborhood_revenue))+
  geom_col(fill="#ffc107")+
  labs(title ="Top 5 neighborhoods in terms of revenue",y="",x="Neighborhood")+
  geom_text(aes(label=scales::comma(neighborhood_revenue,prefix = "Ksh")),color="black",
           hjust=0.6,vjust=0.5,family=f1,size=2.8)+
  coord_flip()+
  custom_theme()+
  theme(
  axis.text.x = element_blank()
  )
```



Q2 - What are the top 5 stores in terms of sales quantity?

Insights:

Milestone Technology Ltd store tops in terms of the sales volume at 26,265 units, followed by Delight Beauty shop (26,110 units), Timko Agency 2 (25,970), Hillocks shop Harambee (25,905 units), Joyous shop (25,767 units).

```r
#filtering top 5 stores in terms of sales quantity
top5_stores <-  sales_revenue_data %>%
  select(shop_location,shop_name,total_sales) %>%
  arrange(desc(total_sales)) %>%
  top_n(5)
```

```
## Selecting by total_sales
```

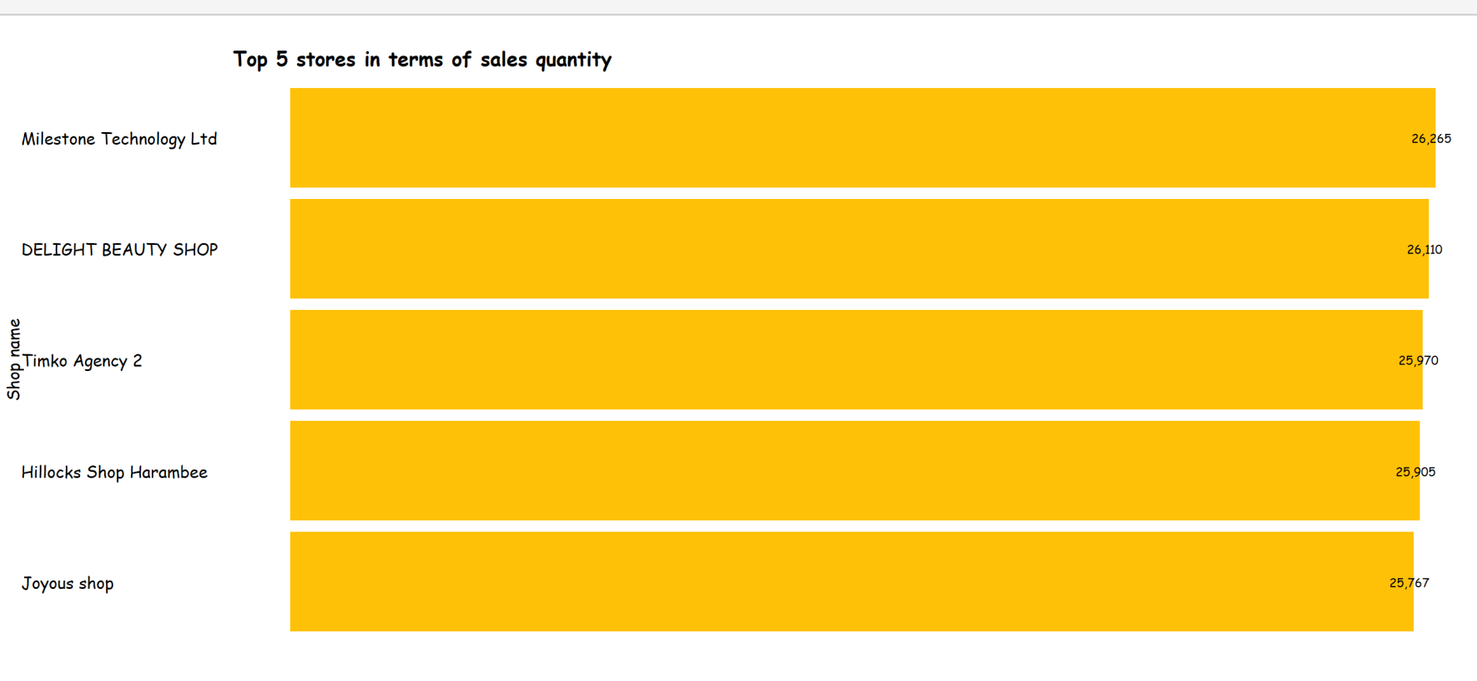
```r
#saving dataset for dashboard
write.csv(top5_stores,"./www/top5_stores_salesqty.csv",row.names = FALSE)

#inspecting data
head(top5_stores)
```

shop_location	shop_name	total_sales
<chr>	<chr>	<dbl>
Huruma	Milestone Technology Ltd	26265
Thika	DELIGHT BEAUTY SHOP	26110
Mwihoko	Timko Agency 2	25970
Buruburu	Hillocks Shop Harambee	25905
Rongai	Joyous shop	25767

5 rows

Plotting top 5 stores in terms of sales quantity

```r
top5_stores %>%
  ggplot(aes(x=reorder(shop_name,total_sales),y=total_sales))+
  geom_col(fill="#ffc107")+
  labs(title ="Top 5 stores in terms of sales quantity",y="",x="Shop name")+
  geom_text(aes(label=scales::comma(total_sales)),color="black",
           hjust=0.6,vjust=0.5,family=f1,size=2.8)+
  coord_flip()+
  custom_theme()+
  theme(
  axis.text.x = element_blank()
  )
```



Q3 - What is the best-selling item in terms of both sales quantity and revenue?

Insights:

Widgets products are the best selling item in terms of sales revenue at Kes 747 509 100. Bobs products are the best selling item in terms of sales volume at 2,592,791 units.

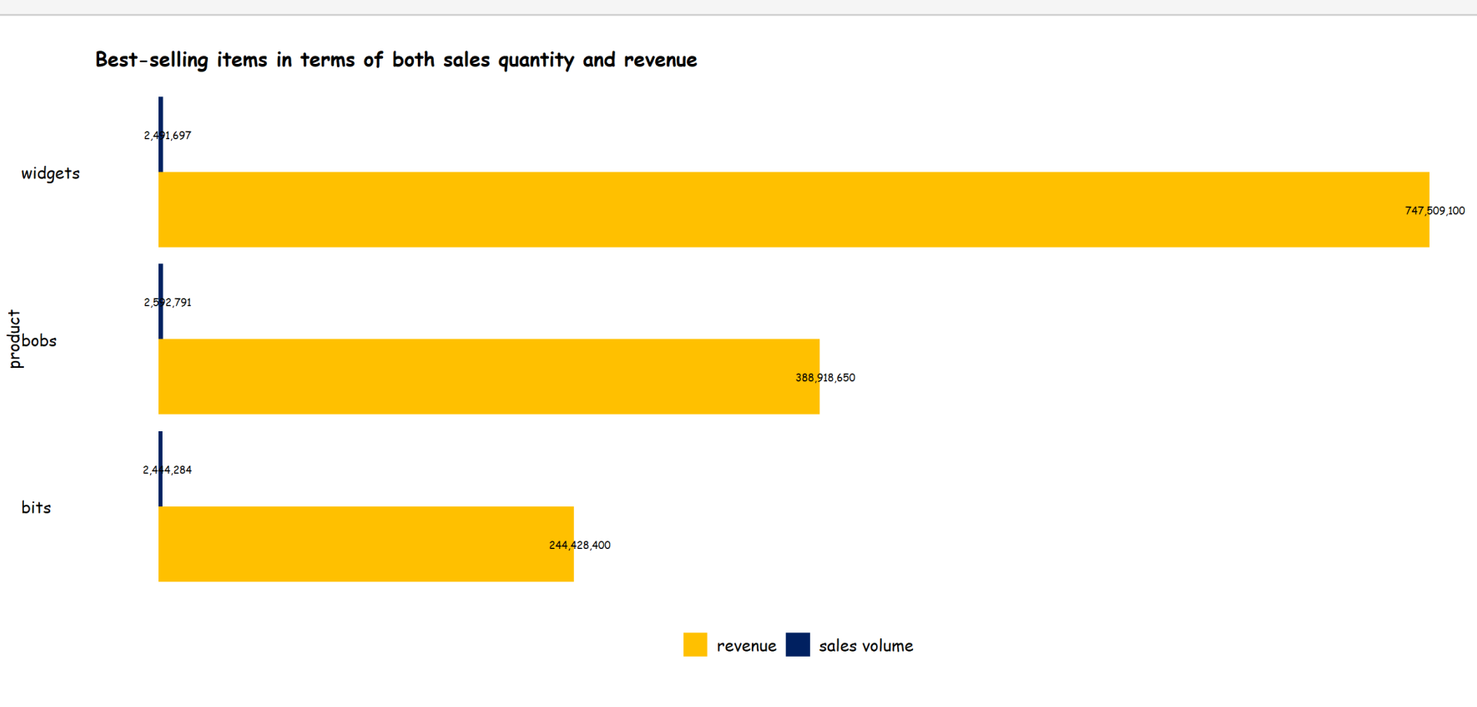
```r
#converting sales_revenue data from wide to long format
#grouping by product to calculate the total sales volume and revenue for each product
sales_revenue_data_long <- sales_revenue_data %>%
  select(bits_qty:widgets_revenue) %>%
  gather(product, amount,bits_qty:widgets_revenue)%>%
  group_by(product) %>%
  summarise(
    total_amount = sum(amount)
  )
sales_revenue_data_long$type <- c("sales volume","revenue","sales volume","revenue","sales volume","revenue")

sales_revenue_data_long$product <-c("bits","bits","bobs","bobs","widgets","widgets")

#saving dataset for dashboard use
write.csv(sales_revenue_data_long,"./www/bestselling_item.csv",row.names = FALSE)

#inspecting the data
head(sales_revenue_data_long)
```

product	total_amount	type
<chr>	<dbl>	<chr>
bits	2444284	sales volume
bits	244428400	revenue
bobs	2592791	sales volume
bobs	388918650	revenue
widgets	2491697	sales volume
widgets	747509100	revenue

6 rows

Plotting the best-selling items in terms of both sales quantity and revenue

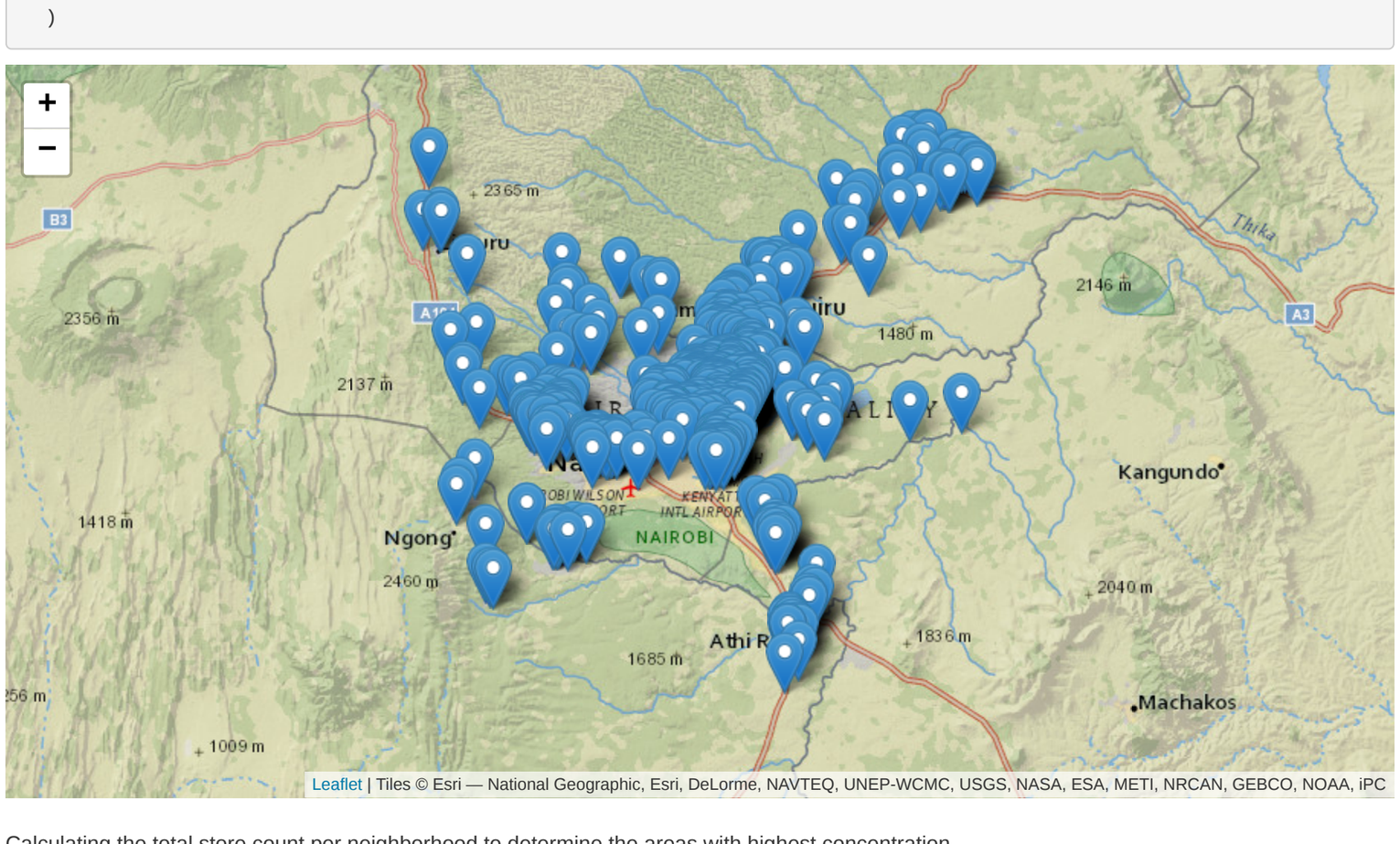
```r
sales_revenue_data_long %>%
  ggplot(aes(x=product,y=total_amount,fill=type))+
  geom_bar(position = "dodge",stat = "identity")+
  geom_text(aes(label=scales::comma(total_amount)),family = f1,size=2.4,hjust=0.4,vjust=0.5,color="black",
            position=position_dodge(width=0.9),face="bold")+
  labs(title = "Best-selling items in terms of both sales quantity and revenue")+
  scale_fill_manual(values=c("#ffc000","#002060"))+
  coord_flip()+
  custom_theme()+
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    legend.title = element_blank(),
    legend.position = "bottom"
  )
```



Q4 - Which part of Nairobi has the highest concentration of stores?

There is a positive correlation between the store concentration and revenue generated. Thika, Kayole, Kibera, Huruma have the highest store concentration and revenue generated as earlier observed.

```r
#making a leaflet map to understand geographical distribution of stores

leaflet(width = "100%") %>%
   # add a default basemap
  addProviderTiles(providers$Esri.NatGeoWorldMap) %>%
  addMarkers(
    data = sales_revenue_data,
    lng = ~lon,
    lat = ~lat,
    # create custom labels
    label = paste(
      "Shop Name:",
      sales_revenue_data$shop_name, "<br>",
      "Shop Location:",
     sales_revenue_data$shop_location, "<br>",
      "Total revenue:",
      sales_revenue_data$total_revenue, "<br>",
      "Total sales quantity: ",
      sales_revenue_data$total_sales
    )
    %>%
      lapply(htmltools::HTML)
  )
```



Calculating the total store count per neighborhood to determine the areas with highest concentration

```r
#calculating the store count
sales_store_count <- sales_revenue_data %>%
  group_by(shop_location) %>%
  count(shop_name) %>%
  summarise(
    shop_count = sum(n)
  )
#sorting the data
top_salesstore_count <- sales_store_count %>%
  arrange(desc(shop_count)) %>% top_n(5)
```

```
## Selecting by shop_count
```

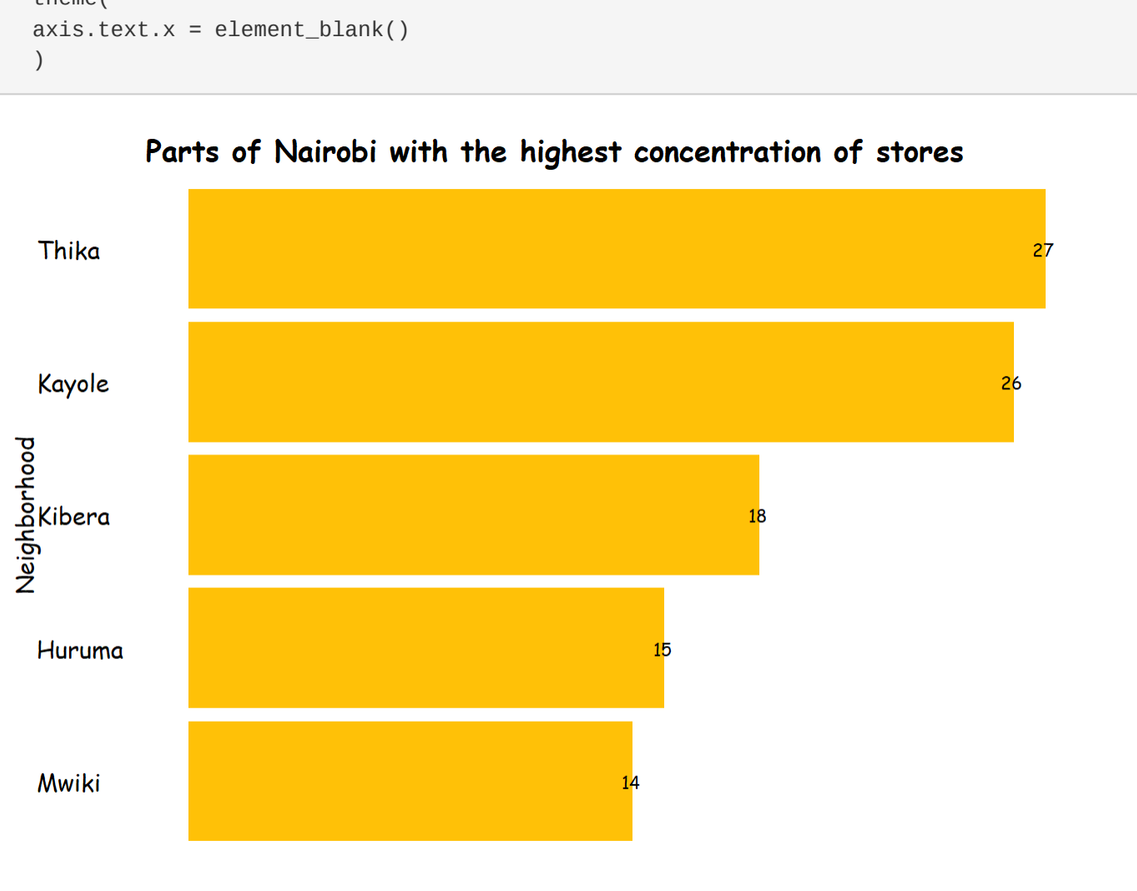
```r
#saving dataset for dashboard use
write.csv(top_salesstore_count,"./www/top_salesstore_count.csv",row.names = FALSE)

#inspecting data
head(top_salesstore_count)
```

shop_location	shop_count
<chr>	<int>
Thika	27
Kayole	26
Kibera	18
Huruma	15
Mwiki	14

5 rows

Plotting areas with the highest concentration of stores

```r
top_salesstore_count %>%  ggplot(aes(x=reorder(shop_location,shop_count),y=shop_count))+
  geom_col(fill="#ffc107")+
  labs(title ="Parts of Nairobi with the highest concentration of stores",y="",x="Neighborhood")+
  geom_text(aes(label=shop_count),color="black",
           hjust=0.6,vjust=0.5,family=f1,size=2.8)+
  coord_flip()+
  custom_theme()+
  theme(
  axis.text.x = element_blank()
  )
```



Part 3 Share - Communicating and interpreting results

We will use R shiny for bringing our data to life by incorporating interactivity. The live dashboard can be accessed here:Krafterkorp Sales Dashboard

Part 4 Act - Putting insights to work

Kraftwerk Korp have insights on store performance, location, sales revenue and volume and will be able to make informed decisions on areas of expansion and other marketing strategies.