COMP4010 - Homework 1

2024-03-10

Homework 1

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
df <- read.csv("ikea_data.csv")
```

Task 1.

```
df$price_usd <- 0.27 * df$price
```

Task 2.

```
df_split <- separate_longer_delim(df, designer, "/")
```

Task 3.

```
top_designers <- df_split %>%
  filter(designer != "IKEA of Sweden" & !is.na(designer)) %>%
  group_by(designer) %>%
  summarise(num_items = n()) %>%
  # slice_max(num_items, n = 20)
  top_n(20, num_items)
sum(df_split$designer == "IKEA of Sweden", na.rm = T)
```

```
## [1] 1502
```

```
sum(is.na(df_split$designer))
```

```
## [1] 143
```

```
sum(df_split$designer == "IKEA of Sweden" | is.na(df_split$designer))
```

```
## [1] 1645
```

Who are the top 3 designers by number of items?

Ehlén Johansson, Francis Cayouette, Ola Wihlborg.

What is the number of items designed by IKEA of Sweden or by unknown designers? Why should we exclude them from the analysis?

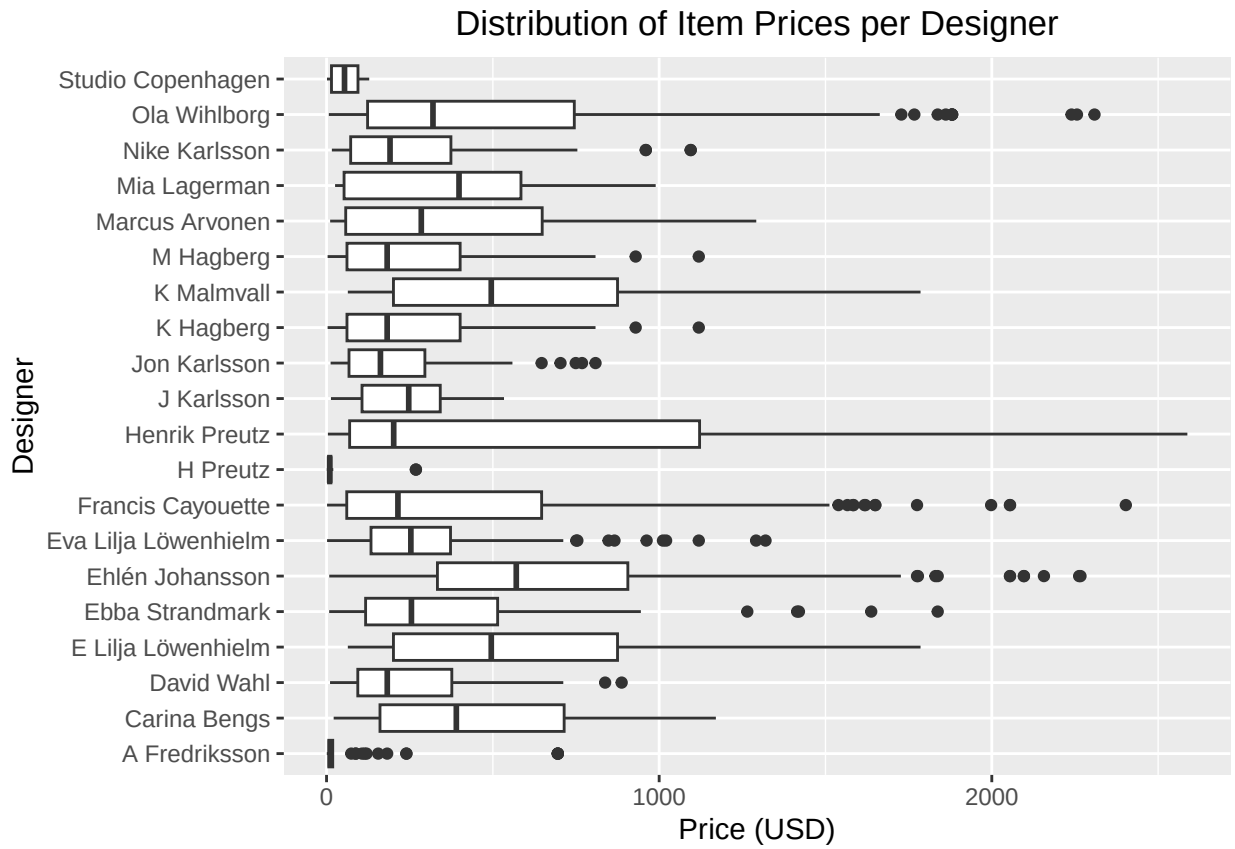
1645. We should exclude them because we're looking to rank designers by the number of products they design. It's not useful to consider products whose designers are unknown or which were designed by the company IKEA of Sweden.

Task 4.

```
designer_filtered_df <- inner_join(df_split, top_designers)
```

```
## Joining with `by = join_by(designer)`
```

```
ggplot(data = designer_filtered_df, aes(x = designer, y = price_usd)) + geom_boxplot() +
coord_flip() +
labs(x = "Designer", y = "Price (USD)",
title = "Distribution of Item Prices per Designer") +
theme(plot.title = element_text(hjust = 0.5))
```



In 3-4 sentences, briefly describe the key findings from this plot. (Open ended question - write down anything useful you can see from this plot)

Overall, 13 out of 20 designers have at least one item priced above $1000, and 4 have at least 1 item priced above $2000

Items designed by Studio Copenhagen and A Fredriksson are all very cheap and close in price. On the other hand, Henrik Preutz designed products of various prices ranges. While the cheaper half of their products have similar prices compared to those of other designers, he is also responsible for the design of the most expensive product.

Also of note is the fact that designers Jon Karlsson and Henrik Preutz also show up in the data as "J Karlsson" and "H Preutz". To obtain more accurate results, further data processing is required to merge cases such as the two mentioned.

In your opinion, is this an effective visualization? If not, what is your suggestion to improve it?

This visualization is not very effective for a number of reasons:

- Multiple name spellings per designer (mentioned above).
- Large value range and extreme outliers (e.g. Henrik Preutz) means that it's hard to distinguish lower value (e.g. Studio Copenhagen and A Fredriksson).

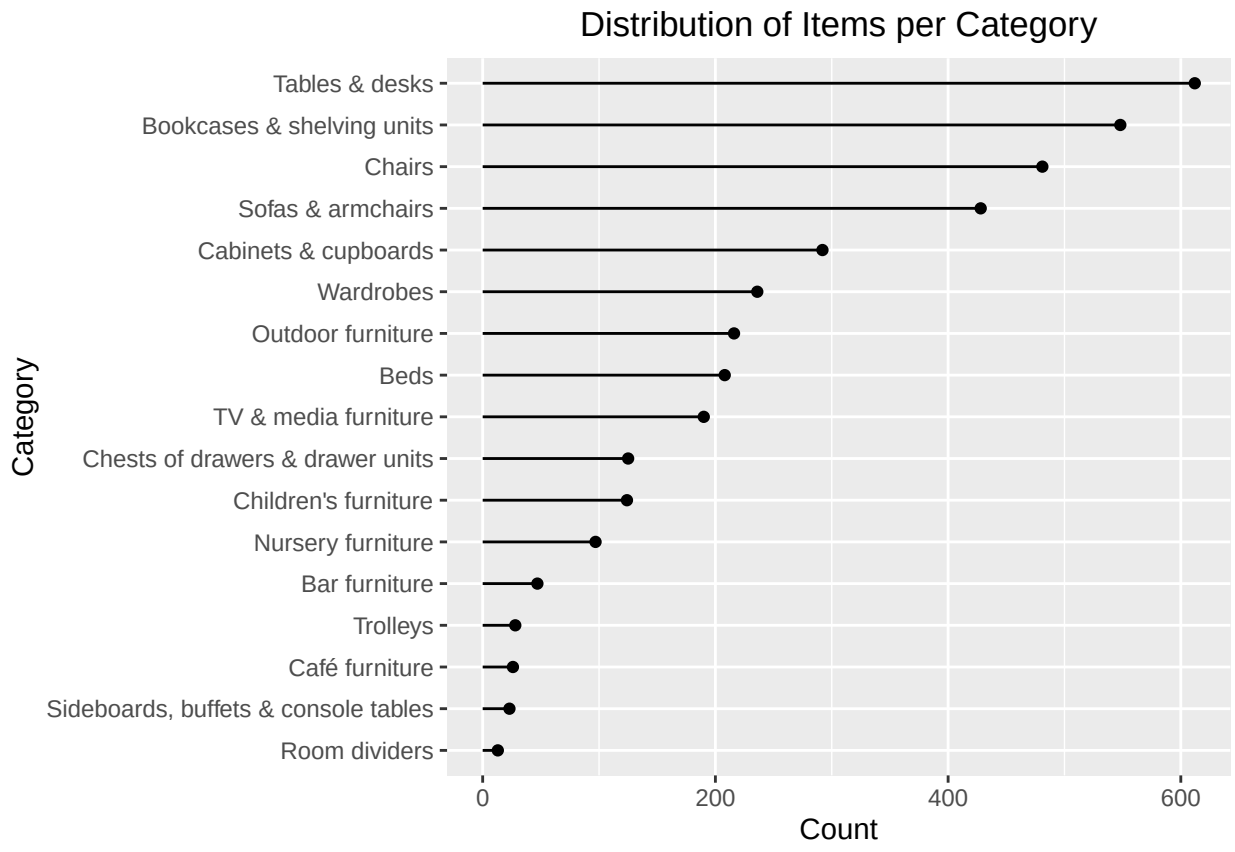
The visualization may be improved by:

- Using a transformed scale (e.g. log scale).
- Processing data.

Task 5.

```
category_df <- df %>%
  group_by(category) %>%
  summarise(num_items = n())


ggplot(
  data = category_df,
  mapping = aes(x = num_items, y = reorder(category, num_items))
) +
  geom_segment(aes(x = 0, xend = num_items, y = category, yend = category)) +
  geom_point() +
  labs(
    x = "Count", y = "Category",
    title = "Distribution of Items per Category"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```



In 3-4 sentences, briefly describe the key findings from this plot.

Tables & desks is the category with the most items, followed by Bookcases & shelving units, Chairs, and Sofas and armchairs. On the other hand, Room dividers is the least diverse category.

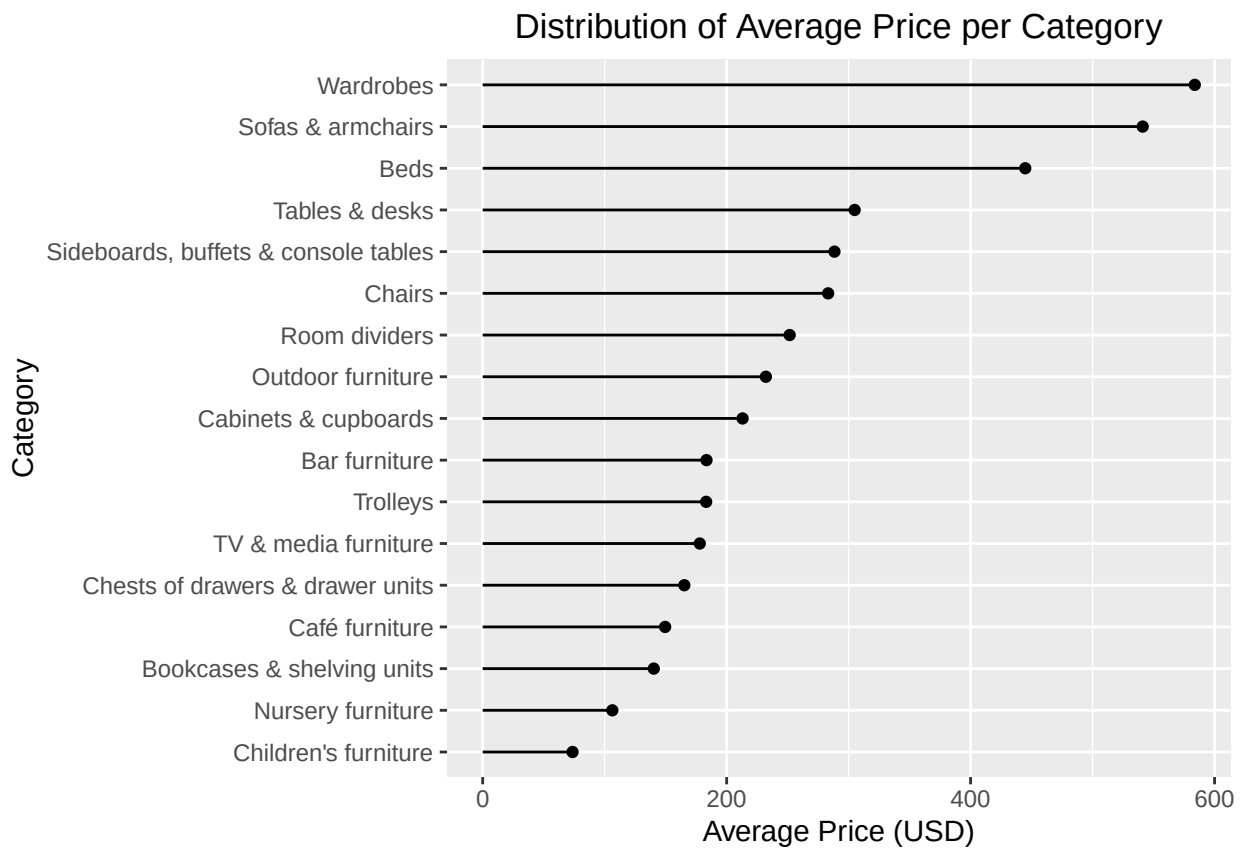
In your opinion, is this an effective visualization? If not, what is your suggestion to improve it?

Yes. It conveys the needed information effectively.

Task 6.

```r
price_by_category_df <- df %>%
  group_by(category) %>%
  summarise(avg_price = mean(price_usd))


ggplot(
  data = price_by_category_df,
  mapping = aes(x = avg_price, y = reorder(category, avg_price))
) +
  geom_segment(aes(x = 0, xend = avg_price, y = category, yend = category)) +
  geom_point() +
  labs(
    x = "Average Price (USD)", y = "Category",
    title = "Distribution of Average Price per Category"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```



In 3-4 sentences, briefly describe the key findings from this plot.

Wardrobes is the category with the highest average cost, followed by Sofas and armchairs, Beds, and Tables & desks. On the other hand, Children's furniture is the cheapest category on average.

In your opinion, is this an effective visualization? If not, what is your suggestion to improve it?

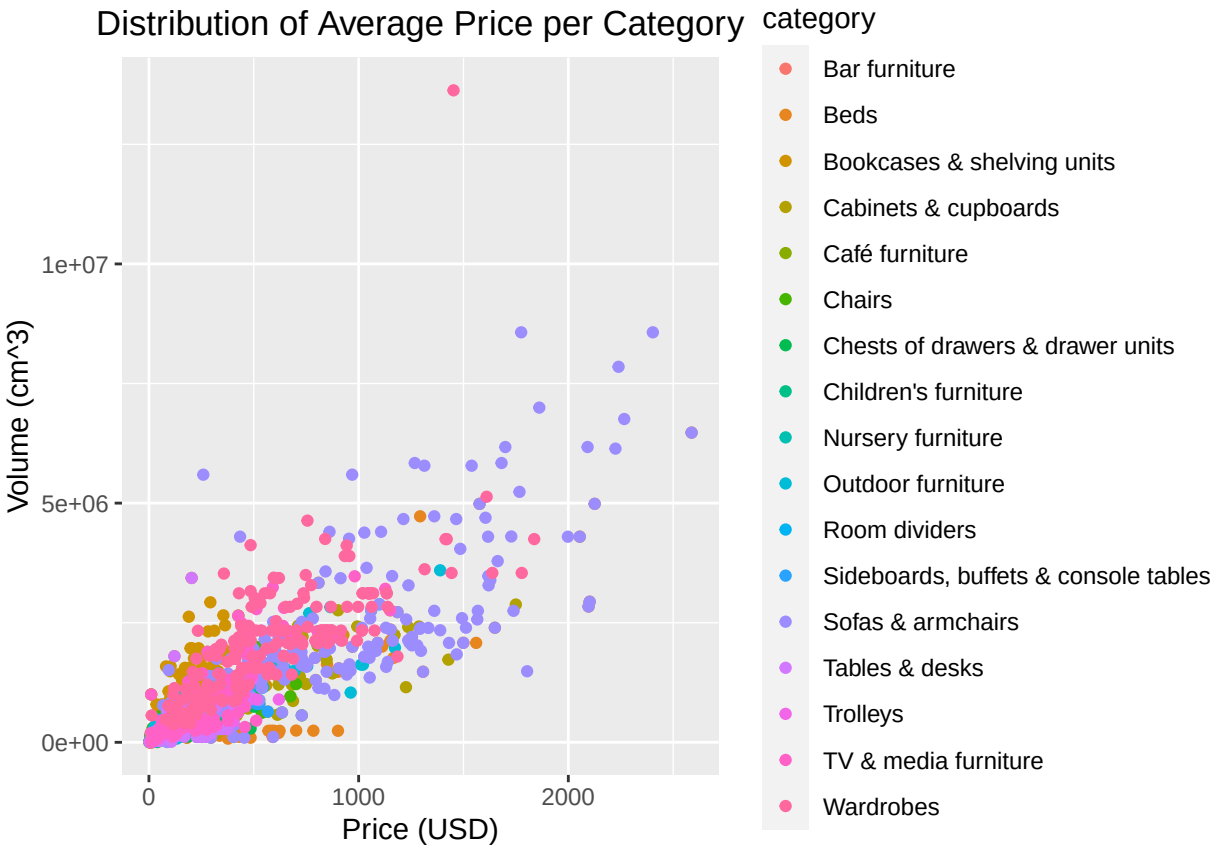
Yes. It conveys the needed information effectively.

Task 7.

```
volume_df <- df %>%
  mutate(volume = depth * height * width)

ggplot(
  data = volume_df,
  mapping = aes(x = price_usd, y = volume, color = category)
) +
  geom_point() +
  labs(
    x = "Price (USD)", y = "Volume (cm^3)",
    title = "Distribution of Average Price per Category"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: Removed 1795 rows containing missing values (`geom_point()`).
```



In 3-4 sentences, briefly describe the key findings from this plot.

There seems to be a correlation between volume and price. Generally, Sofas and armchairs are some of the biggest **and** most expensive items, with one exception: a wardrobe item is by far the largest item in terms of volume; however, it is not the most expensive item.

In your opinion, is this an effective visualization? If not, what is your suggestion to improve it?

While the visualization effectively conveys certain information, most of the points are clustered in a corner making it hard to distinguish between points.

This visualization may be improved by:

- Reducing the marker size and increasing transparency.
- Transforming the scale to accommodate outliers