Data-Story

Setup

```
library(imager)
library(wordcloud2)
library(tidyverse)
library(dplyr)
library(plotly)

data <- data.frame(read.csv('../data/cleaned-womens-shoe-prices.csv'))
```

How many shoes are there in the dataset?

```
nrow(data)
```

```
## [1] 4555
```

How many brands?

```
data %>%
  select(brand) %>%
  unique() %>%
  nrow
```

```
## [1] 967
```

What are these brands?

Make a wordcloud to have a brief overview. The bigger the word is, the more shoes the brand has in this dataset.

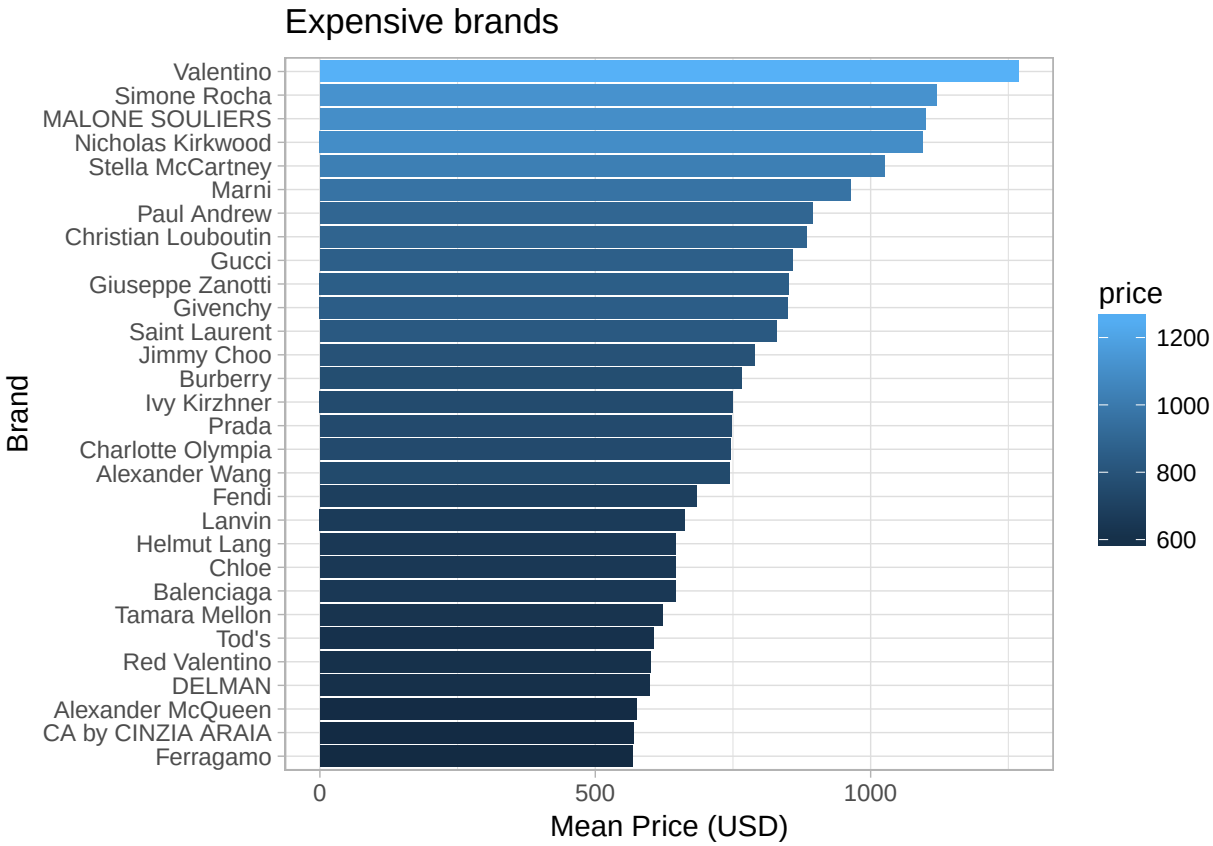
```
df.wc <- data %>%
  select(brand, price.avg) %>%
  group_by(brand) %>%
  dplyr::summarise(n = n()) %>%
  arrange(desc(n))

wordcloud2(df.wc)
```

We can see some familiar names, such as the huge Nine West, Puma, etc, though there appears to be just a few UGG shoes in this dataset.

What are the expensive brands?

```
data %>%
  group_by(brand) %>%
  dplyr::summarise(price = mean(price.avg, rm.na=true)) %>%
  filter(price > 100) %>%
  arrange(desc(price)) %>%
  top_n(30) %>%
  ggplot(mapping = aes(x=reorder(brand, price), y=price)) +
    geom_bar(stat = "identity", aes(fill=price)) +
    theme_light() +
    coord_flip() +
    labs(title="Expensive brands", x="Brand", y="Mean Price (USD)")
```

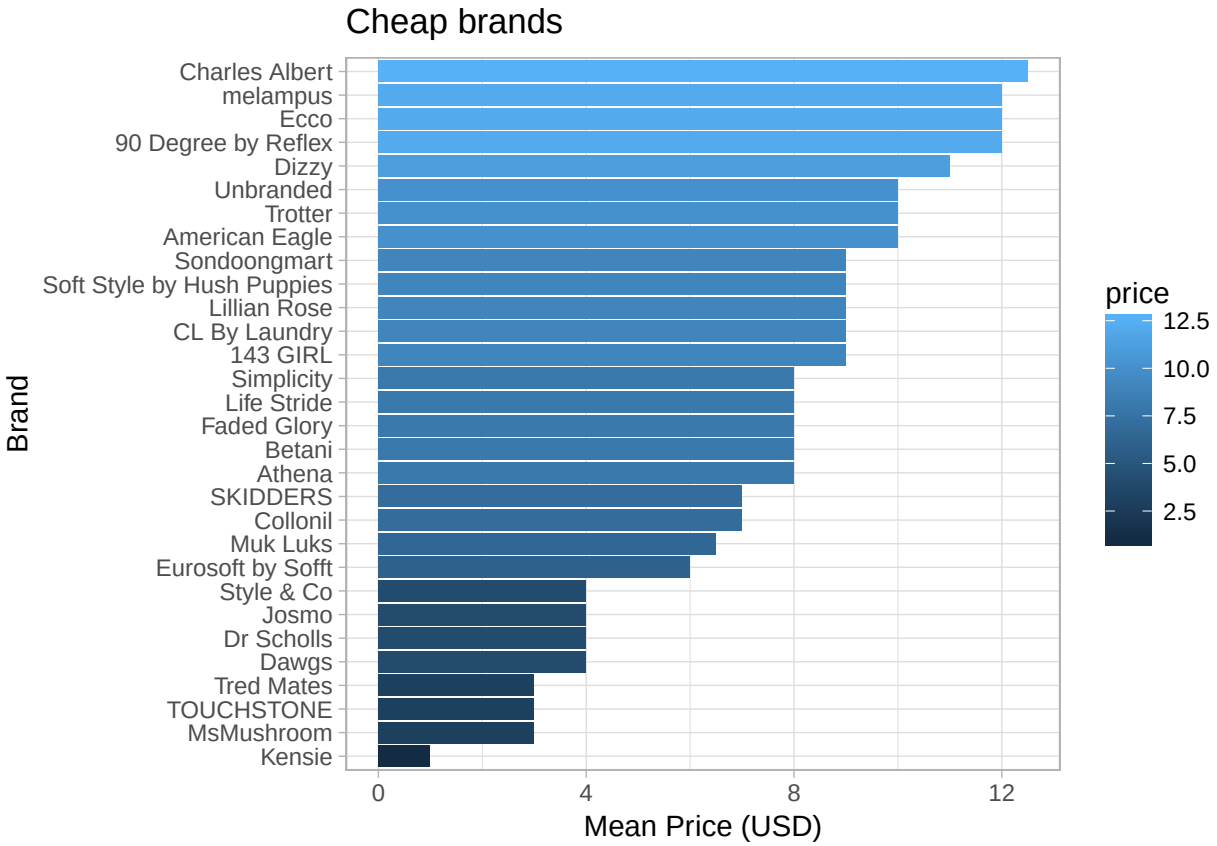


Top 5 are over $1000, and top 40 are over $500.

How about the cheaper ones?

```
data %>%
  group_by(brand) %>%
  dplyr::summarise(price = mean(price.avg, rm.na=true)) %>%
  filter(price < 50) %>%
  arrange(desc(price)) %>%
  top_n(-30) %>%
  ggplot(mapping = aes(x=reorder(brand, price), y=price)) +
    geom_bar(stat = "identity", aes(fill=price)) +
    theme_light() +
    scale_colour_gradient() +
    coord_flip() +
    labs(title="Cheap brands", x="Brand", y="Mean Price (USD)")
```

```
## Selecting by price
```

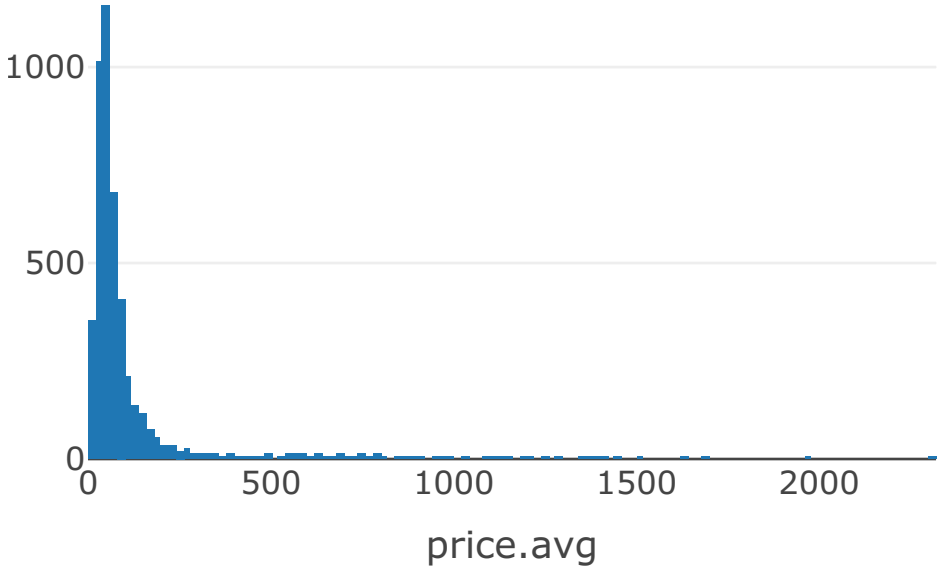


Kensie has $1 shoes!!!

We just saw the mean price for each brand, but how about the prices for individual shoes?

Plot the distribution of the shoe prices in this dataset.

```
plot_ly(data, x = ~price.avg, type = "histogram")
```

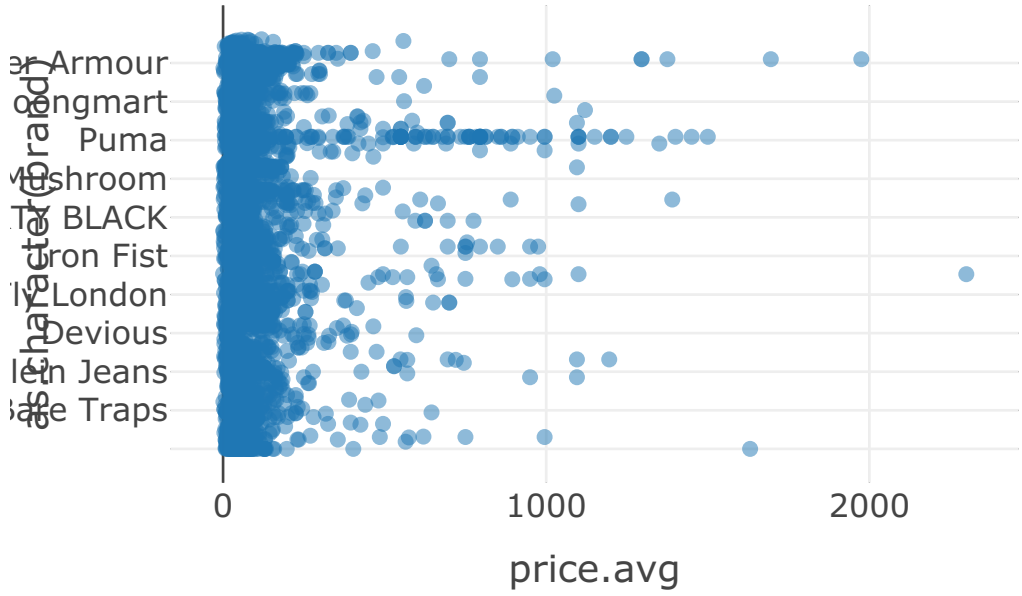


There is a pair of $2000+ shoes. What does it look like? It will be figured out as we further explore the distributions for the brands, rather than the mean prices.

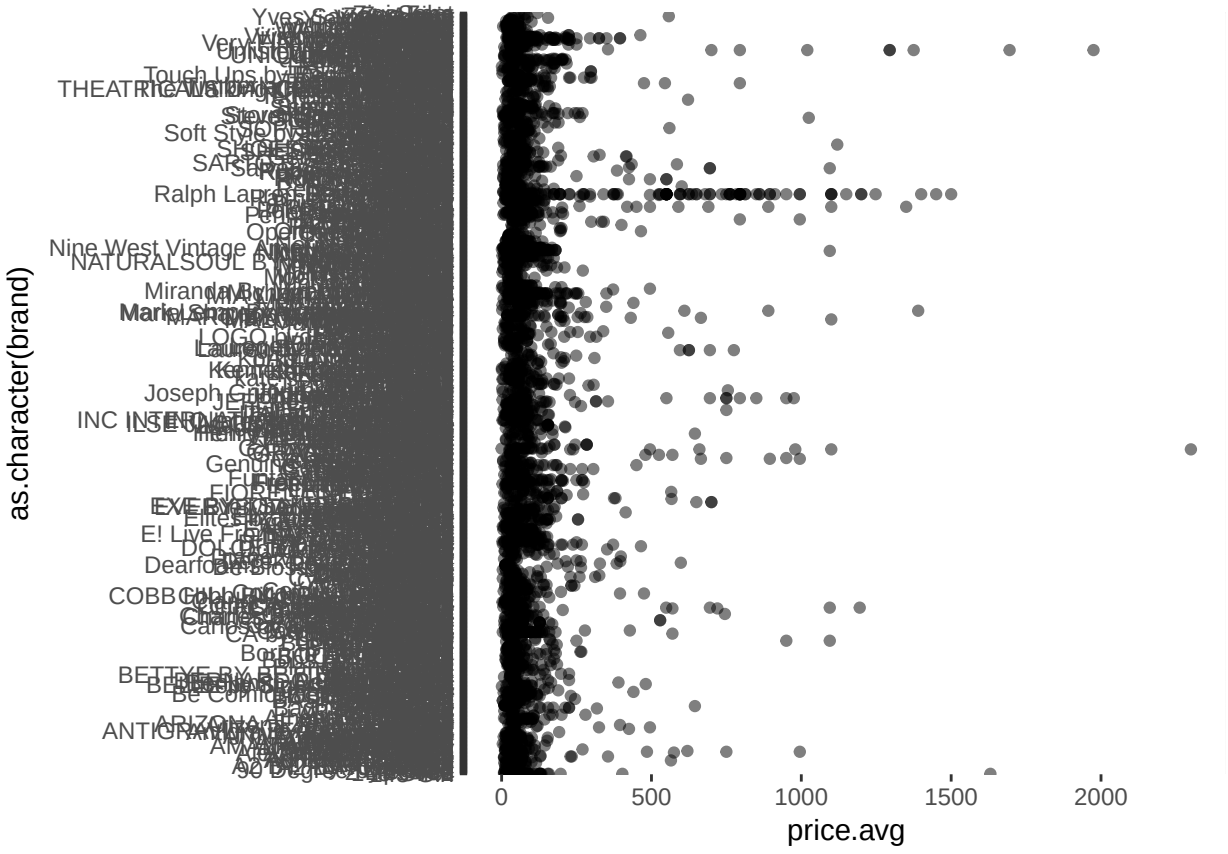
Price distribution for each brand?

```
data %>%
  plot_ly(x = ~price.avg, y = ~as.character(brand), type = "scatter", alpha = 0.5)
```

```
## No scatter mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plot.ly/r/reference/#scatter-mode
```




```
data %>%
  ggplot(aes(x=price.avg, y=as.character(brand))) + geom_point(alpha=0.5)
```



(Note: Originally plotly was used to create interactive visualizations, but the charts are not showing in html, thus later ggplot was used instead.)

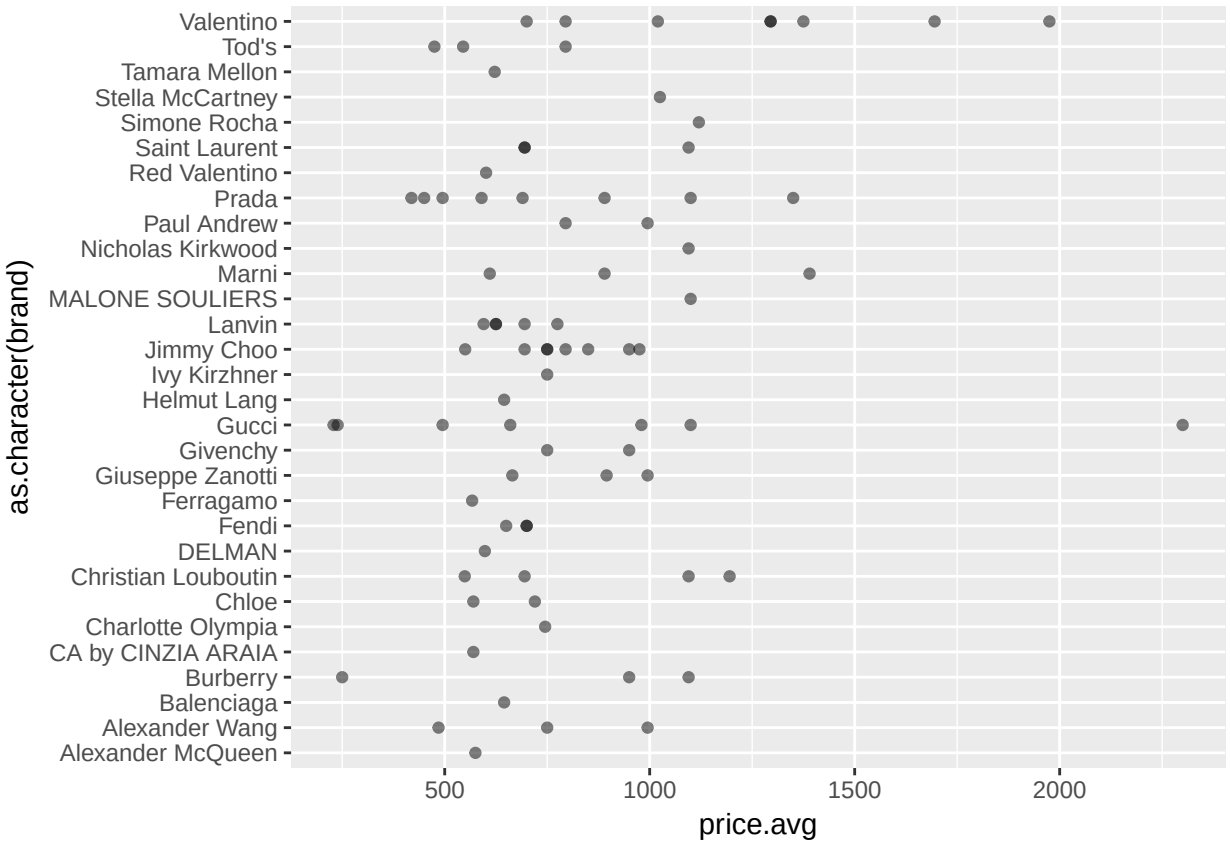
The $2000+ shoes are Gucci. By zooming in and out, some other facts can be revealed, e.g., the prices of Ralph Lauren's shoes are very spread-out.

Expensive brands prices?

```
expensive_brands <- data %>%
  group_by(brand) %>%
  dplyr::summarise(price = mean(price.avg, rm.na=true)) %>%
  filter(price > 100) %>%
  arrange(desc(price)) %>%
  top_n(30)
```

```
## Selecting by price
```

```
data %>%
  filter(brand %in% expensive_brands$brand) %>%
  ggplot(aes(x=price.avg, y=as.character(brand))) + geom_point(alpha=0.5)
```



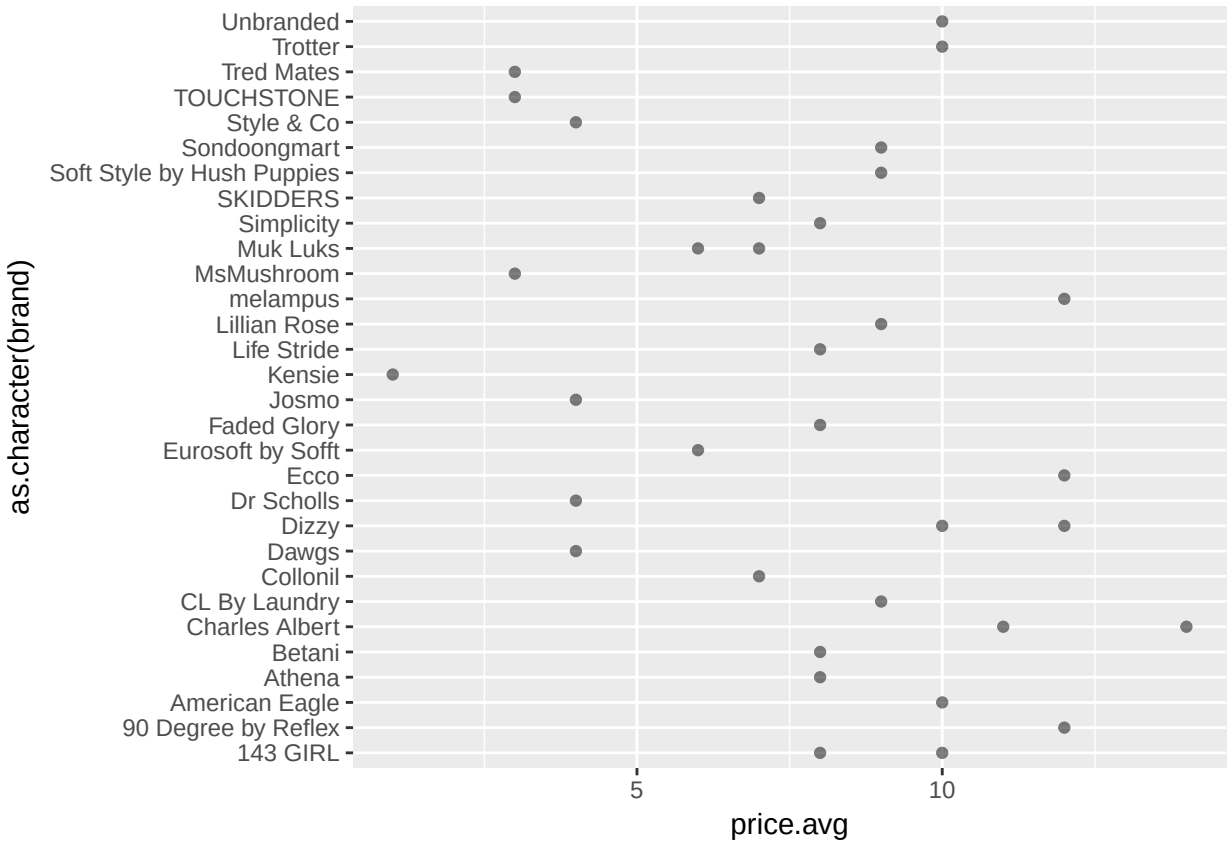
The top brands are indeed expensive, of which the prices range from $230 to $2300.

Expensive brands prices?

```
cheap_brands <- data %>%
  group_by(brand) %>%
  dplyr::summarise(price = mean(price.avg, rm.na=true)) %>%
  filter(price < 100) %>%
  arrange(desc(price)) %>%
  top_n(-30)
```

```
## Selecting by price
```

```
data %>%
  filter(brand %in% cheap_brands$brand) %>%
  ggplot(aes(x=price.avg, y=as.character(brand))) + geom_point(alpha=0.5)
```



Cheap brands's prices range from $1 to $14.

What do these expensive or cheap shoes look like?

An interactive visualization based on D3 was made to better display shoes details, including images, color choices, etc. The intended purpose of the d3 vis is not to show a discovered trend, but to support (maybe can also encourage) a user to further explore the dataset.