Preliminary data wrangling

Marta Sapizhak

```r
raw_data <- read_csv("shopping_trends.csv")
head(raw_data) |>
  kable() |>
  kable_styling(bootstrap_options = c("striped"))
```

Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size
1	55	Male	Blouse	Clothing	53	Kentucky	L
2	19	Male	Sweater	Clothing	64	Maine	L
3	50	Male	Jeans	Clothing	73	Massachusetts	S
4	21	Male	Sandals	Footwear	90	Rhode Island	M
5	45	Male	Blouse	Clothing	49	Oregon	M
6	46	Male	Sneakers	Footwear	20	Wyoming	M

There are no missing values and the dataset is overall cleaned. Some of the things we need to account for to make sure we compare apples to apples are:

- gender - one of key explanatory
- items and categories - one of key explanatory - need to group better to lower the number of categ variables
- size - the key explanatory
- prices - response variable? how to measure the chances that the same thing for a man will be cheaper? Multiple regression?
- season - one of key explanatory
- age - may be one of the further analysis explanatory.

```r
nrow_before_clean <- (nrow(raw_data))

# Remove specified columns
cols_to_remove <- c("Frequency of Purchases", "Previous Purchases",
                    "Preferred Payment Method", "Payment Method",
```

```r
                   "Review Rating", "Discount Applied", "Promo Code Used",
                   "Subscription Status", "Shipping Type")

data <- raw_data |>
  filter(`Discount Applied` == 'No', `Promo Code Used` == "No") |>
  select(-all_of(cols_to_remove))

nrow_after_clean <- (nrow(data))
```

Our initial data has had 3900 observations initially and after the filtering, it's 2223

The sales taxes correlation was basically not significant and small enough to disregard. Similarly since gender will be similarly distributed across values, we can disregard differences in shipping prices and other things, since the information about it is limited and we expect them to be fairly even between the two genders considered in the study.

```r
# Read tax data
tax_data <- read.csv("LOST_July_2024_Rate_Table.csv")

# Clean tax data - remove % and convert to numeric
tax_data$Combined_Rate <- as.numeric(sub("%", "", tax_data$Combined.Rate))

# Calculate median prices by state
state_prices <- data |>
  group_by(Location) |>
  summarise(
    median_price = median(`Purchase Amount (USD)`),
    n_transactions = n()
  )

# Merge with tax data
comparison <- merge(state_prices, tax_data,
                   by.x="Location", by.y="State")

# Run linear regression
model <- lm(median_price ~ Combined_Rate, data=comparison)

print(tidy(model))
```

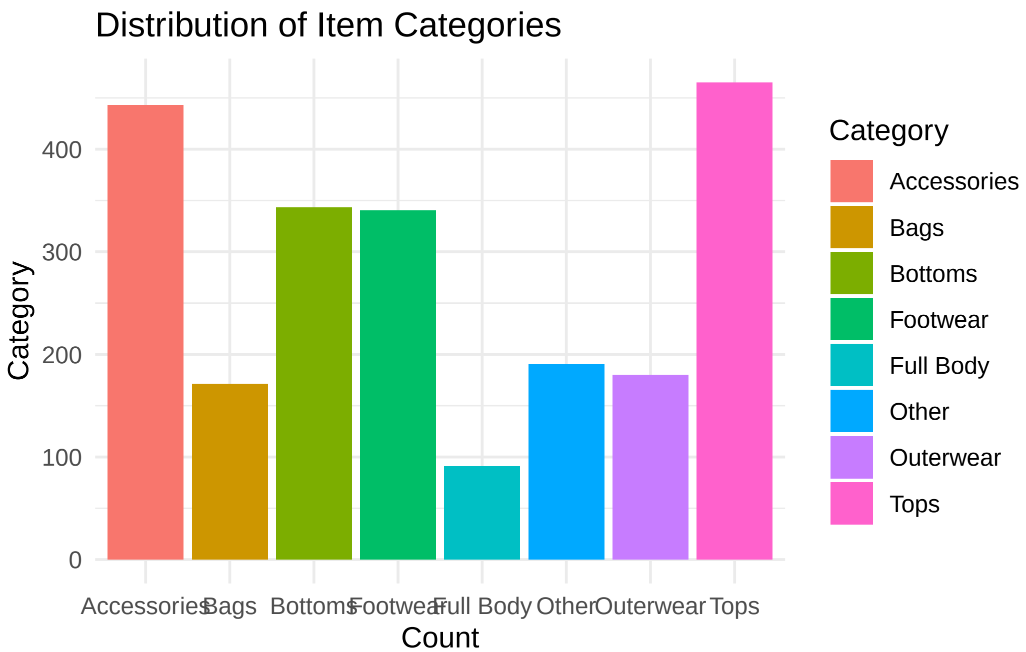
```
# A tibble: 2 x 5
  term          estimate std.error statistic  p.value
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>
```

```
1 (Intercept)      57.2       3.04      18.8   1.83e-21
2 Combined_Rate     0.366     0.424      0.864 3.93e- 1
```

One thing we definitely want to account for is the different item types. To simplify the analysis we want to separate them into more categories than we are given, since those aren't thoroughly informative but at the same time we need less unique values for this categorical variable.

```r
# First, let's see all unique items
unique_items <- sort(unique(data$`Item.Purchased`))

# Create a function to categorize items
categorize_items <- function(item) {
  tops <- c("T-shirt", "Blouse", "Shirt", "Tank Top", "Sweater", "Hoodie", "Sweatshirt", "Ca
  bottoms <- c("Pants", "Jeans", "Shorts", "Skirt", "Leggings", "Trousers")
  full_body <- c("Dress", "Suit", "Jumpsuit", "Romper")
  outerwear <- c("Jacket", "Coat", "Blazer", "Windbreaker")
  accessories <- c("Belt", "Scarf", "Hat", "Cap", "Gloves", "Tie", "Socks")
  footwear <- c("Shoes", "Boots", "Sandals", "Sneakers", "Heels")
  bags <- c("Backpack", "Handbag", "Purse", "Wallet", "Tote")

  # Categorize
  if(item %in% tops) return("Tops")
  if(item %in% bottoms) return("Bottoms")
  if(item %in% full_body) return("Full Body")
  if(item %in% outerwear) return("Outerwear")
  if(item %in% accessories) return("Accessories")
  if(item %in% footwear) return("Footwear")
  if(item %in% bags) return("Bags")
  return("Other")
}

# Apply categorization
data$Item_Category <- sapply(data$`Item Purchased`, categorize_items)

# Get summary of new categories
category_summary <- table(data$Item_Category)
category_df <- data.frame(
  Category = names(category_summary),
  Count = as.numeric(category_summary),
  Percentage = round(as.numeric(prop.table(category_summary)) * 100, 2)
)
```

```r
ggplot(category_df, aes(x = Count, y = Category, fill = Category)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Distribution of Item Categories",
       x = "Category",
       y = "Count")
```



Does season is what I think it is - the season of the item style or is it when it was purchased? One thing that came up - idk why sometimes i need to call it as Price.Paid vs `Price Paid`

```r
head(data |>
  filter(Category == "Outerwear")) |>
  kable() |>
  kable_styling(bootstrap_options = c("striped"))
```

Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	S
1684	23	Male	Coat	Outerwear	57	Arkansas	S
1691	40	Male	Coat	Outerwear	30	Nevada	M
1694	63	Male	Jacket	Outerwear	88	Tennessee	X
1714	49	Male	Jacket	Outerwear	22	Missouri	M

	1727	57	Male	Coat	Outerwear	28	Indiana
	1748	61	Male	Coat	Outerwear	23	New Hampshire

```r
# Analyze by Season
season_item_analysis <- data |>
 group_by(Season, `Item Purchased`) |>
 summarise(
   count = n(),
   .groups = 'drop'
 ) |>
 arrange(Season, desc(count))

# Get top items for each season
top_items_by_season <- season_item_analysis |>
 group_by(Season) |>
 slice_max(order_by = count, n = 3)

top_items_by_season |>
  kable() |>
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Season	Item Purchased	count
Fall	Socks	33
Fall	Skirt	31
Fall	Handbag	30
Fall	Jacket	30
Spring	Sandals	32
Spring	Sweater	32
Spring	Blouse	30
Summer	Blouse	31
Summer	Dress	28
Summer	Jewelry	28
Summer	Scarf	28
Summer	Socks	28
Winter	Sunglasses	32
Winter	Shirt	31
Winter	Socks	28

Seems like the season might be impacting data not in the way that we'd expect. For now, just still treat season as an explanatory variable.

Now that we have all settled - look at the data again:

```r
data
```

```
# A tibble: 2,223 x 11
   `Customer ID`   Age Gender `Item Purchased` Category   Purchase Amount (USD~1
           <dbl> <dbl> <chr>  <chr>            <chr>                       <dbl>
 1          1678    65 Male   Jeans            Clothing                       35
 2          1679    41 Male   Pants            Clothing                       71
 3          1680    60 Male   Dress            Clothing                       52
 4          1681    61 Male   Shoes            Footwear                       37
 5          1682    24 Male   Sneakers         Footwear                       95
 6          1683    65 Male   Socks            Clothing                       97
 7          1684    23 Male   Coat             Outerwear                      57
 8          1685    30 Male   Shirt            Clothing                       93
 9          1686    33 Male   Blouse           Clothing                       48
10          1687    22 Male   Gloves           Accessori~                     75
# i 2,213 more rows
# i abbreviated name: 1: `Purchase Amount (USD)`
# i 5 more variables: Location <chr>, Size <chr>, Color <chr>, Season <chr>,
#   Item_Category <chr>
```

```r
table(data$Size)
```

```

   L    M    S   XL 
 596 1013  362  252 
```

Our primary question is whether the gender is very important for when buying XL, L clothes. To account for confounding variables - the model explanatory are: item types, season, location, age, color(pink tax?) specifically for pink??

- prices - response variable? how to measure the chances that the same thing for a man will be cheaper? Multiple regression?
- season - one of key explanatory
- age - may be one of the further analysis explanatory.

At this stage there are two questions we can pose - /

Multiple linear regression modeling:/ Do women pay more for larger size of clothing? - think about what categories and items it involves

What variable is most significant in determining the gender of the buyer?