Informative Title

Ava Cascario, Sydney Holt, and Kacie Rohn

2024-12-03

```
# Load Packages
library(ggplot2)
library(dplyr)

shoppingTrendsRaw <- read.csv(
  file = "shopping_trends.csv",
  header = TRUE,
  sep = ","
)
typeReviews <- shoppingTrendsRaw %>%
  select(
    Gender, Review.Rating
  ) %>%
  group_by(
    Gender, Review.Rating
  )
ggplot(
  data = typeReviews,
  aes(
    x = Gender,
    y = Review.Rating
  )
) +
  geom_boxplot()
shoppingSummary <- lm(formula = Purchase.Amount..USD. ~ Category, data = shoppingTrendsRaw)
summary(shoppingSummary)
ageGenderItems <- shoppingTrendsRaw %>%
  select(
    Gender, Age, Item.Purchased
  ) %>%
  group_by(
    Age, Gender, Item.Purchased
  ) %>%
  summarize(
    Count = n(),
  )
```

```
ggplot(
  data = ageGenderItems,
  aes(
    x = Item.Purchased,
    y = Count,
    fill = Gender
  )
) +
  geom_bar(
    stat = "identity",
    position = position_dodge(width = 5)
  ) +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1)
  )
itemPurchasedData <- shoppingTrendsRaw %>%
  group_by(
    Location,
    Season,
    Item.Purchased
  ) %>%
  summarize(
    Item.Count = n(),
    .groups = "drop"
  )
stateToRegion <- c(
  "Maine" = "Northeast",
  "New Hampshire" = "Northeast",
  "Vermont" = "Northeast",
  "Massachusetts" = "Northeast",
  "Rhode Island" = "Northeast",
  "Connecticut" = "Northeast",
  "New York" = "Northeast",
  "New Jersey" = "Northeast",
  "Pennsylvania" = "Northeast",
  "Delaware" = "South",
  "Maryland" = "South",
  "Virginia" = "South",
  "North Carolina" = "South",
  "South Carolina" = "South",
  "Georgia" = "South",
  "Florida" = "South",
  "West Virginia" = "South",
  "Kentucky" = "South",
  "Tennessee" = "South",
  "Alabama" = "South",
  "Mississippi" = "South",
```

```
  "Arkansas" = "South",
  "Louisiana" = "South",
  "Oklahoma" = "South",
  "Texas" = "South",
  "Indiana" = "Midwest",
  "Illinois" = "Midwest",
  "Michigan" = "Midwest",
  "Ohio" = "Midwest",
  "Wisconsin" = "Midwest",
  "Missouri" = "Midwest",
  "Iowa" = "Midwest",
  "Minnesota" = "Midwest",
  "North Dakota" = "Midwest",
  "South Dakota" = "Midwest",
  "Nebraska" = "Midwest",
  "Kansas" = "Midwest",
  "Montana" = "West",
  "Wyoming" = "West",
  "Colorado" = "West",
  "Idaho" = "West",
  "Nevada" = "West",
  "Utah" = "West",
  "Arizona" = "West",
  "New Mexico" = "West",
  "Washington" = "West",
  "Oregon" = "West",
  "California" = "West",
  "Alaska" = "West",
  "Hawaii" = "West"
)

itemPurchasedData <- itemPurchasedData %>%
  mutate(
    Region = stateToRegion[Location]
  )
ggplot(
  itemPurchasedData,
  aes(
    x = Item.Purchased,
    y = Item.Count,
    fill = Region
  )
) +
  geom_bar(
    stat = "Identity",
    position = "dodge"
  ) +
```

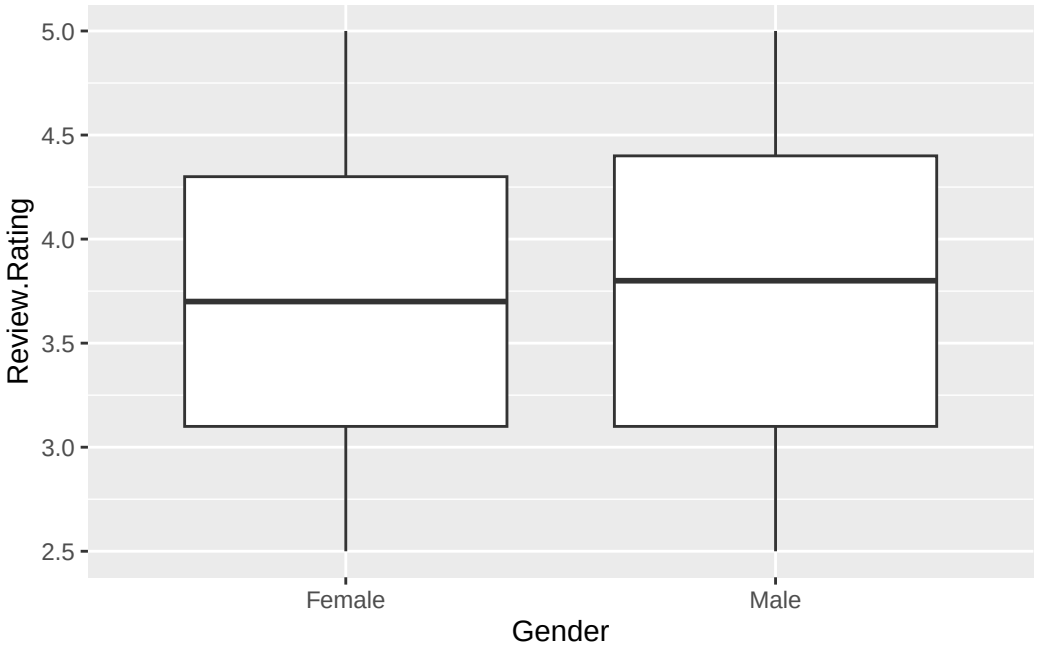
```
  facet_wrap(
    ~Season,
    scales = "free_x",
    ncol = 1,
    strip.position = "top"
  ) +
  labs(
    title = "Items Bought by Region and Season",
    x = "Item",
    y = "Number of Items Bought",
    fill = "Region"
  ) +
  scale_fill_manual(
    values = c(
      "Northeast" = "#779ECB",
      "South" = "#486856",
      "Midwest" = "#FFD580",
      "West" = "#FF7F7F"
    )
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(
      angle = 45,
      hjust = 1
    ),
    legend.position = "right",
    strip.text = element_text(size = 12),
    plot.title = element_text(
      size = 18,
      hjust = 0.5,
      face = "bold"
    ),
    plot.margin = margin(20, 20, 20, 20),
    axis.title.y = element_text(
      margin = margin(r = 10)
    )
  )
```

Research Topic: Field of Study vs Occupation

Research Questions

Provenance Of Our Data

FAIR and CARE Principles

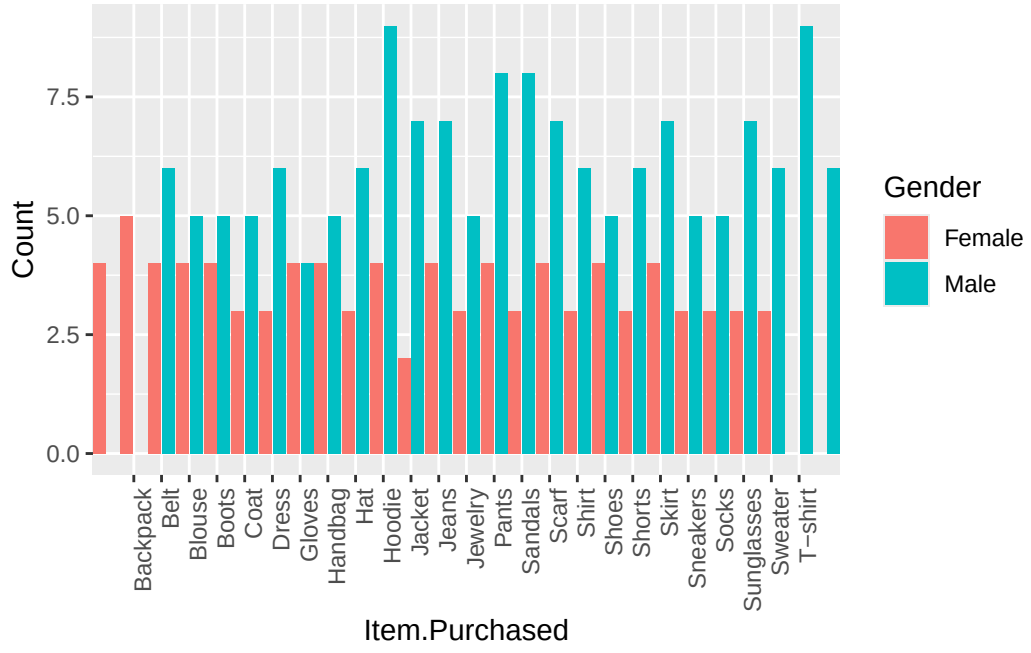



```
Call:
lm(formula = Purchase.Amount..USD. ~ Category, data = shoppingTrendsRaw)

Residuals:
    Min      1Q  Median      3Q     Max
-40.255 -21.025  -0.025  20.975  42.827

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        59.8387     0.6725  88.979   <2e-16 ***
CategoryClothing    0.1866     0.8804   0.212   0.8321
CategoryFootwear    0.4167     1.1783   0.354   0.7236
CategoryOuterwear  -2.6659     1.4775  -1.804   0.0713 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 23.68 on 3896 degrees of freedom
Multiple R-squared:  0.001118,  Adjusted R-squared:  0.0003489
F-statistic: 1.454 on 3 and 3896 DF,  p-value: 0.2252
```



Research Topic: Field of Study vs Occupation

Our focus in conducting our study is to look at the association between field of study and actual job occupation. That is, to know if what an individual studied in college would have an affect on the job they get post-graduation. We will conduct our research by looking at different research questions and creating visualizations to represent the data that correlates with each question. We will then explore our outcomes and summarize our findings and how they connect back to our topic.

Research Questions

The first question we will explore is, how does your field of study correlate with your job occupation? We want to know what kind of association is found between the two factors, and if what an individual studied has an affect on the job they pursue. We would also like to explore, what the most commonly chosen occupations are based on the field of study? For example, if an individual was a software engineering major, what is the most commonly chosen job after graduation. Finally, we would like to explore the question, is there an association between gender and field of occupation? Not only will our research focus on field of study and occupation, but we would also like to analyze whether gender plays a role in this as well.

Provenance Of Our Data

We are utilizing a data set that we found on Kaggle. Kaggle is a website focused towards data scientists with a goal in helping others learn about data. The data was collected by Jahnavi Paliwal, a data science masters student at the University of San Francisco. The data has a usability score of 10.0 and is annually updated by Jahnvai, with the latest update being one month ago. This means the data is collected from a reliable source, and updated consistently to remain relevant. The data set is designed to help you predict whether individuals are likely to change their occupation based on their academic background, job experience, and other demographic factors. This can help you answer questions based on numerous aspects of the job industry including human resources, income, industry analysis, job markets, and job availability. This data set contains over 30,000 records each with 22 attributes. This data set constitutes a case as a single individual.

FAIR and CARE Principles