# DATA 606 Data Project Proposal

Kory Martin

March 21, 2023

## Contents

### Introduction

This project will attempt to build a multiple-linear regression model that allows us to predict the total spend for a user based on their gender, age group, department and their payment type. Although this specific data set is related to a series of physical retail shopping malls in Turkey, the general concept can be applied to an e-commerce store or other digital retail experience. The idea is that if we know some information about the profile of our customers - either by them logging into the site or using other digital signals, then we can anticipate their spending and then use this to both predict our sales revenue for a particular day, and this can be used as a basis for a machine learning algorithm that can help us anticipate if we are on track to meet our daily sales goals and determine if we need to dynamically generate special promotions in order to induce the customer to make addtional purchases, etc.

### Data Preparation

```r
library(tidyverse)
library(kableExtra)
library(forcats)

path <-  '/Users/korymartin/Library/Mobile Documents/com~apple~CloudDocs/Grad Programs/CUNY SPS/Final P

customer_df <- read.csv(path)
customer_df <- tibble(customer_df)
customer_df <- janitor::clean_names(customer_df)
```

### Research question

The basic research question is can we build a multiple-linear regression model that can be used to predict customer sales for a transaction, given their gender, age group, category of shopping and payment method.

### Cases

For the dataset that I'm using, there are 99,457 cases. Each case represents a single categorical transaction by a customer.

### Data collection

For this project, I'll be using the **Customer Shopping Dataset - Retail Sales Data** downloaded from Kaggle

### Type of study

This will be an observational study. I will be analyzing the data collected for the various schools in the city pertaining to their performance in the most recent school year that the data is available.

### Data Source

https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset?resource=download

### Dependent Variable

The dependent variable for this analysis is price, which represents the total spend by a single customer within a specific product category.

### Independent Variable(s)

At the onset, the indepdendent variables that we expect to use in this analysis are: - gender - age (which will be binned) - product_category - payment_method

It's possible that we will incorporate some additional synthetic variables into the model based on exploratory analysis.

### Relevant summary statistics

```r
# Bin Age
customer_df <- customer_df %>%
  mutate(age_bin = case_when(
    age < 20 ~ 'under_20',
    between(age, 20,29) ~ '20_to_29',
    between(age, 30,39) ~ '30_to_39',
    between(age, 40,49) ~ '40_to_49',
    between(age, 50,59) ~ '50_to_59',
    between(age, 60,69) ~ '60_to_69',
    TRUE ~ 'other'
  ))

customer_df <- customer_df %>%
  mutate(avg_price = price/quantity)
```

**Modify Data  Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.**

Table 1: Number of Transactions by Category

| category | n |
|---|---:|
| Clothing | 34487 |
| Cosmetics | 15097 |
| Food & Beverage | 14776 |
| Toys | 10087 |
| Shoes | 10034 |
| Souvenir | 4999 |
| Technology | 4996 |
| Books | 4981 |

Table 2: Number of Transactions by Gender

| gender | n |
|---|---:|
| Female | 59482 |
| Male | 39975 |

- Show breakdown of transaction data based on gender
- Show breakdown of transaction data based on age_bin
- Show summary statistics for entire dataset (age, gender, category, payment type)

### Overall Data Summary

Here's a breakdown of the overall counts of data (i.e. transactions) broken down by product category, gender, age_bin and payment_method.

```
customer_df %>%
  count(category, sort = T) %>%
  kable(
    caption = 'Number of Transactions by Category'
  ) %>%
  kable_material(c("striped"))
```

```
customer_df %>%
  count(gender, sort=T) %>%
  kable(
    caption = 'Number of Transactions by Gender'
  ) %>%
  kable_material(c("striped"))
```

```
customer_df %>%
  count(payment_method, sort = T) %>%
  kable(
    caption = 'Number of Transactions by Payment Method'
  ) %>%
  kable_material(c("striped"))
```

Table 3: Number of Transactions by Payment Method

| payment_method | n |
|---|---:|
| Cash | 44447 |
| Credit Card | 34931 |
| Debit Card | 20079 |

Table 4: Number of Transactions by Age Bin

| age_bin | n |
|---|---:|
| 30_to_39 | 19287 |
| 20_to_29 | 19263 |
| 40_to_49 | 19153 |
| 60_to_69 | 19043 |
| 50_to_59 | 18931 |
| under_20 | 3780 |

```
customer_df %>%
  count(age_bin, sort = T) %>%
  kable(
    caption = 'Number of Transactions by Age Bin'
  ) %>%
  kable_material(c("striped"))
```

### Gender Data

Summary statistics based on the gender of the customer

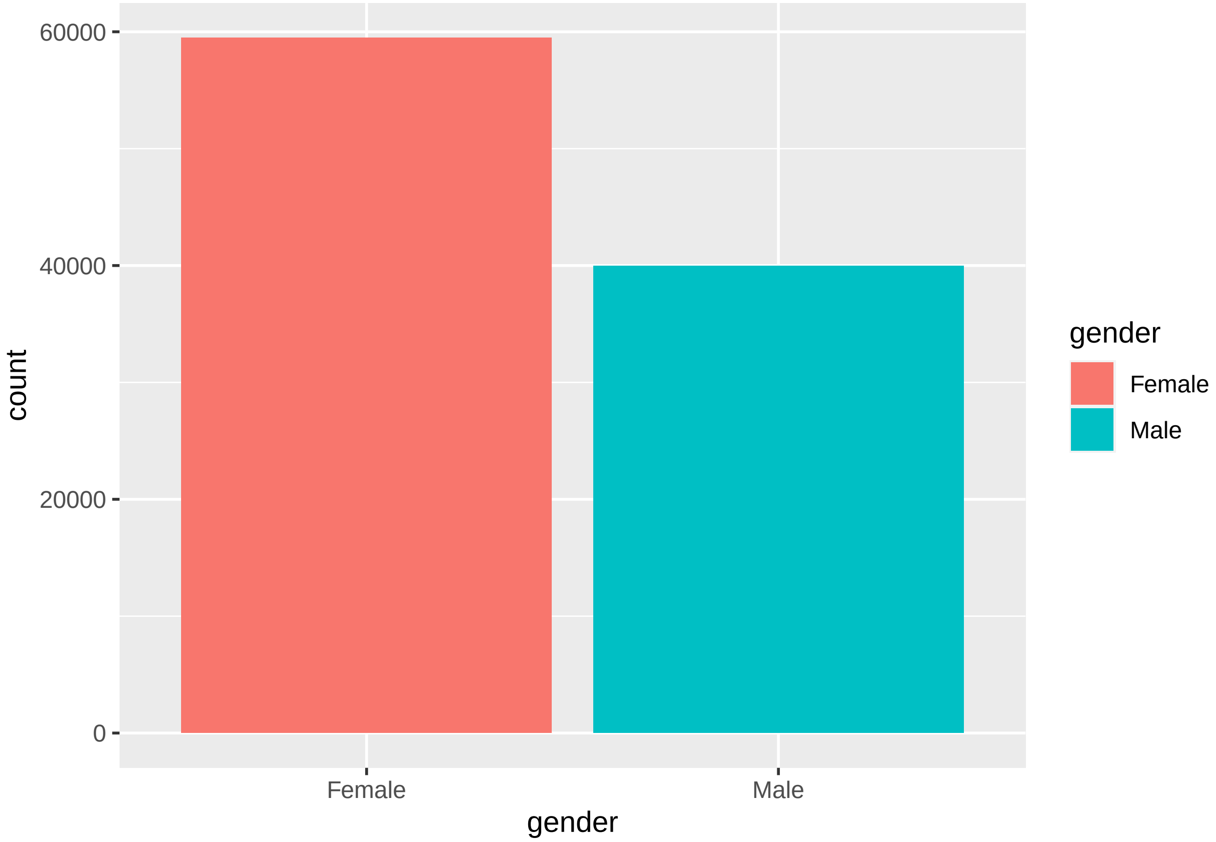
```
## Gender Data
gender_summary <- customer_df %>%
  group_by(gender) %>%
  summarize(n = n(),
            num_txns = sum(quantity),
            avg_age = mean(age),
            median_age = median(age),
            sd_age = sd(age),
            min_age = min(age),
            max_age = max(age),
            total_spend = sum(price),
            mean_price = mean(price),
            median_price = median(price),
            sd_price = sd(price))

gender_summary %>%
  mutate(n = sprintf(n, fmt='%#.0i'),
         num_txns = sprintf(num_txns, fmt='%#.2f'),
         avg_age = sprintf(avg_age, fmt='%#.2f'),
         median_age = sprintf(median_age, fmt='%#.2f'),
         sd_age = sprintf(sd_age, fmt='%#.2f'),
         min_age = sprintf(min_age, fmt='%#.2f'),
         max_age = sprintf(max_age, fmt='%#.2f'),
         total_spend = sprintf(total_spend, fmt='%#.2f'),
```

| stat_name | Female | Male |
|---|---|---|
| n | 59482 | 39975 |
| num_txns | 178659.00 | 120053.00 |
| avg_age | 43.45 | 43.39 |
| median_age | 43.00 | 43.00 |
| sd_age | 14.97 | 15.03 |
| min_age | 18.00 | 18.00 |
| max_age | 69.00 | 69.00 |
| total_spend | 40931801.62 | 27619564.29 |
| mean_price | 688.14 | 690.92 |
| median_price | 203.30 | 203.30 |
| sd_price | 940.79 | 941.78 |

```
         mean_price = sprintf(mean_price, fmt='%#.2f'),
         median_price = sprintf(median_price, fmt='%#.2f'),
         sd_price = sprintf(sd_price, fmt='%#.2f')) %>%
  pivot_longer(!gender, names_to = 'stat_name', values_to = 'stat') %>%
  pivot_wider(names_from = gender, values_from = stat) %>%
  kable() %>%
  kable_material(c("striped"))
```

```
ggplot(customer_df, aes(x=gender, fill=gender)) +
  geom_bar()
```



Category Data

Table 5: Summary Metrics based on Product Category

| category | num_txns | total_spend | pct_txns |
|---|---:|---:|---:|
| Books | 14982 | 226977.3 | 0.0501553 |
| Clothing | 103558 | 31075684.6 | 0.3466818 |
| Cosmetics | 45465 | 1848606.9 | 0.1522035 |
| Food & Beverage | 44277 | 231568.7 | 0.1482264 |
| Shoes | 30217 | 18135336.9 | 0.1011576 |
| Souvenir | 14871 | 174436.8 | 0.0497837 |
| Technology | 15021 | 15772050.0 | 0.0502859 |
| Toys | 30321 | 1086704.6 | 0.1015058 |

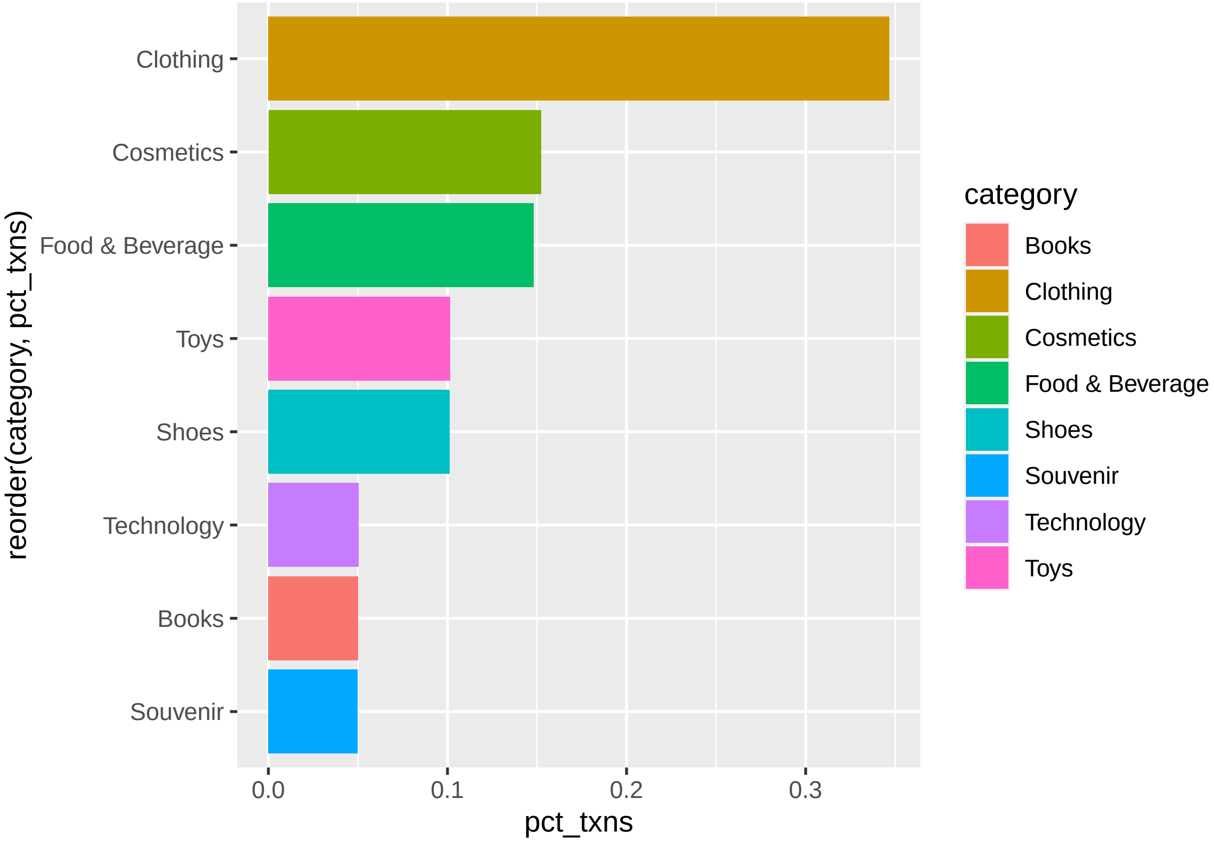
```r
## Category Data
category_summary <- customer_df %>%
  group_by(category) %>%
  summarize(num_txns = sum(quantity),
            total_spend = sum(price)) %>%
  mutate(pct_txns = num_txns/sum(num_txns))


category_summary %>%
  kable(
    caption = "Summary Metrics based on Product Category"
  ) %>%
  kable_material(c("striped"))
```

```r
ggplot(category_summary, aes(x=reorder(category, pct_txns), y=pct_txns, fill=category)) +
  geom_bar(stat='identity') +
  coord_flip()
```

Table 6: Summary Metrics based on Age-Bin

| age_bin | n | total_spend | num_txns | avg_spend | median_spend |
|---|---|---|---|---|---|
| 20_to_29 | 19263 | 13324658 | 57949 | 691.7229 | 203.30 |
| 30_to_39 | 19287 | 13245828 | 57875 | 686.7749 | 203.30 |
| 40_to_49 | 19153 | 13376514 | 57517 | 698.4031 | 300.08 |
| 50_to_59 | 18931 | 12937020 | 56922 | 683.3775 | 203.30 |
| 60_to_69 | 19043 | 13160725 | 57153 | 691.1057 | 203.30 |
| under_20 | 3780 | 2506620 | 11296 | 663.1270 | 203.30 |



Age Data

```
## Age Data

customer_df %>%
  group_by(age_bin) %>%
  summarize(n = n(),
            total_spend = sum(price),
            num_txns = sum(quantity),
            avg_spend = mean(price),
            median_spend = median(price)) %>%
  kable(
    caption = "Summary Metrics based on Age-Bin"
  ) %>%
  kable_material(c("striped"))
```

Table 7: Summary Metrics based on Payment Method

| payment_method | n | total_spend | num_txns | avg_spend | median_spend |
|---|--:|--:|--:|--:|--:|
| Cash | 44447 | 30705031 | 133370 | 690.8235 | 203.3 |
| Credit Card | 34931 | 24051477 | 105045 | 688.5425 | 203.3 |
| Debit Card | 20079 | 13794858 | 60297 | 687.0291 | 203.3 |

Payment Method

```r
customer_df %>%
  group_by(payment_method) %>%
  summarize(n = n(),
            total_spend = sum(price),
            num_txns = sum(quantity),
            avg_spend = mean(price),
            median_spend = median(price)) %>%
  kable(
    caption = "Summary Metrics based on Payment Method"
  ) %>%
  kable_material(c("striped"))
```

### Notable Observations

One thing that I notice is that the dataset is structured where several aspects of the data appear to be pretty balanced and not reflective of the type of distribution one would expect to occur naturally (e.g. avg spend, median spend, distribution of age). I'm not sure how this will affect the potential model, but I think that I will need to use synthetic data to change the distribution to a normal distribution.