Customer Segmentation EDA Report

2023-07-29

Overview

This report aims to explore the customer segmentation dataset and gain insights into different features it contains. It has 6718 lines of non-null data in total. I conduct EDA analysis using tidyverse(Wickham et al. 2019) and ggplot2(Wickham 2016). Please note that the data set is for practice purposes only and, therefore, does not necessarily reflect reality.

First, the below table provides a snippet of the first a few lines of data. The data set consists of 8 features, `Gender`, `Ever_Married`, `Age`, `Graduated`, `Profession`, `Work_Experience`, `Spending_Score` and `Family_Size`.

```
kable(head(clean_data)) |>
  kable_styling()
```

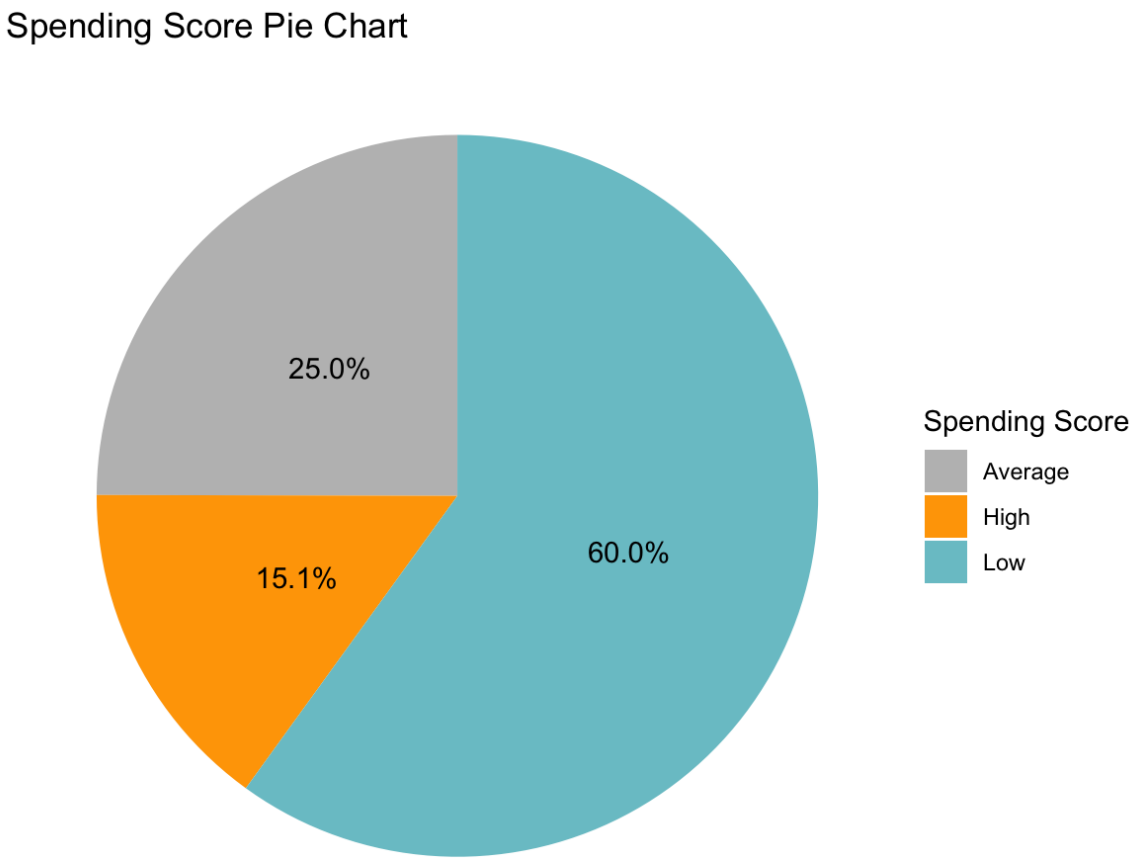
Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size
Male	No	22	No	Healthcare	1	Low	4
Female	Yes	67	Yes	Engineer	1	Low	1
Male	Yes	67	Yes	Lawyer	0	High	2
Male	Yes	56	No	Artist	0	Average	2
Male	No	32	Yes	Healthcare	1	Low	3
Female	No	33	Yes	Healthcare	1	Low	3

Exploratory Data Analysis (EDA)

Spending Score Pie Chart

From the Spending Score Pie Chart, we can see that more than half of the customers in the dataset have a low spending score.

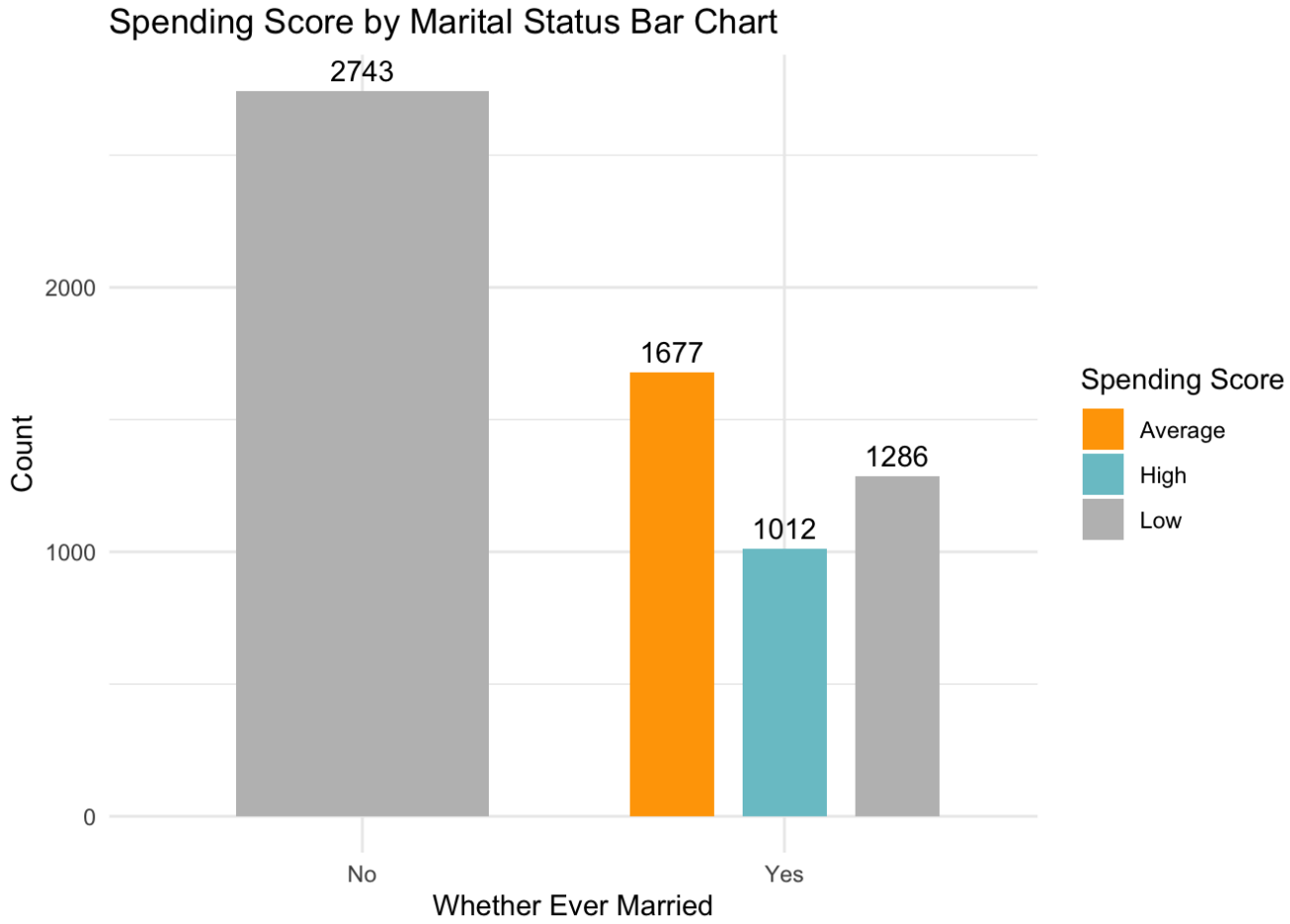
```
spending_score_count <- clean_data |>
  group_by(Spending_Score) |>
  summarize(count = n()) |>
  mutate(percentage = count / sum(count))

spending_score_pie_chart <- ggplot(spending_score_count, aes(x = "", y = count, fill = Sp
ending_Score)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  theme_void() +
  labs(title = "Spending Score Pie Chart", fill = "Spending Score") +
  geom_text(aes(label = scales::percent(percentage)), position = position_stack(vjust =
0.5)) +
  scale_fill_manual(values = c("Average" = "gray", "High" = "orange", "Low" = "cadetblue
3"))
spending_score_pie_chart
```



Spending Score Count by Marital Status Bar Chart

It is intriguing to observe that all unmarried customers exhibit a low spending score, whereas married customers demonstrate a more diverse range of spending scores, including high, low, or average spending scores.

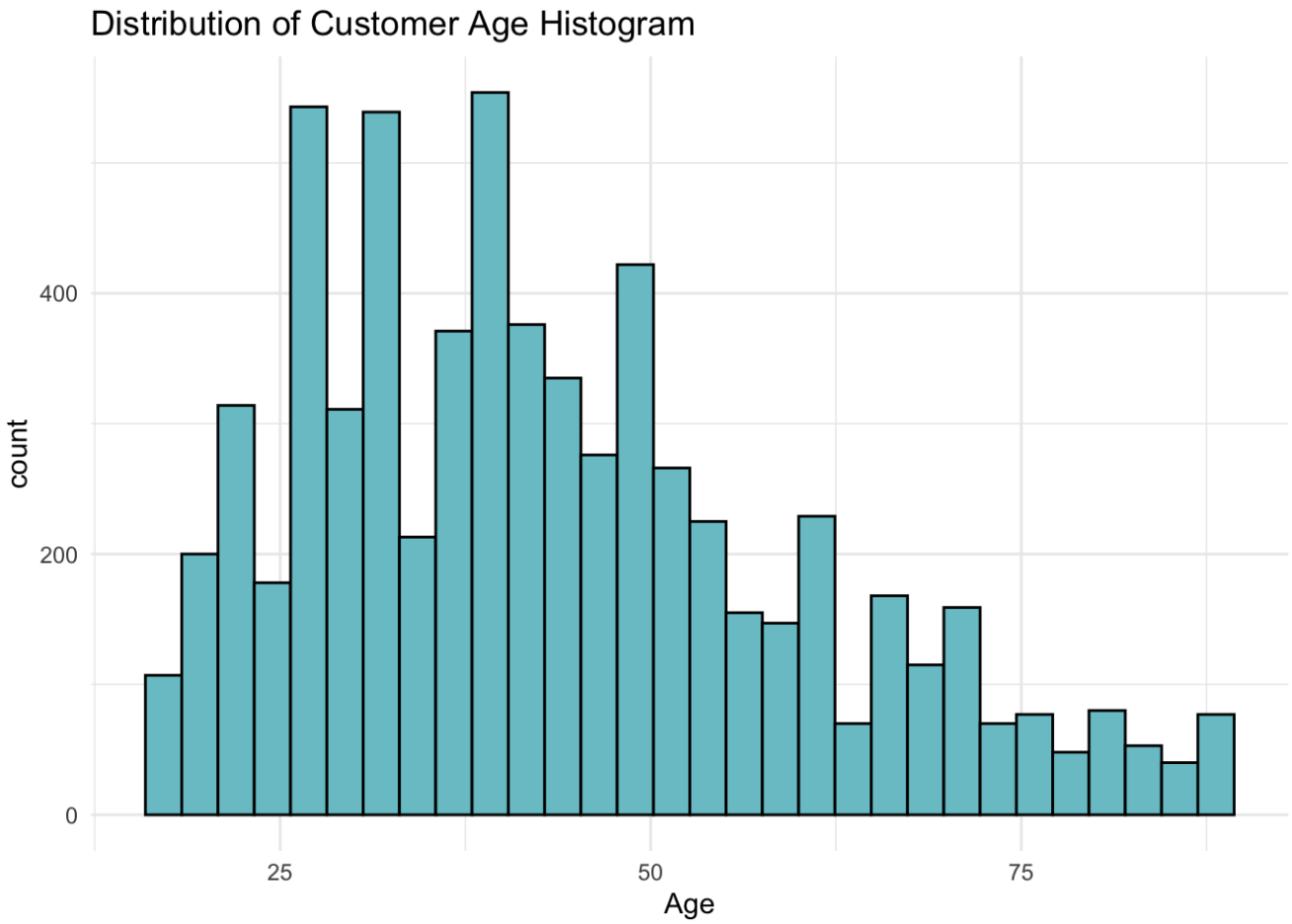
```
spending_score_martial_bar_chart <- clean_data |>
  group_by(Ever_Married, Spending_Score) |>
  summarize(count = n(), .groups = "drop") |>
  ggplot(aes(x = Ever_Married, y = count, fill = Spending_Score)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.6) +
  scale_fill_manual(values = c("High" = "cadetblue3", "Average" = "orange", "Low" = "gra
y")) +
  labs(title = "Spending Score by Marital Status Bar Chart", x = "Whether Ever Married",
y = "Count", fill = "Spending Score") +
  geom_text(aes(label = count), position = position_dodge(width = 0.8), vjust = -0.5)
spending_score_martial_bar_chart
```



Customer Age Distribution Histogram

In the dataset, there are more customers under 50 years old than over 50 years old.

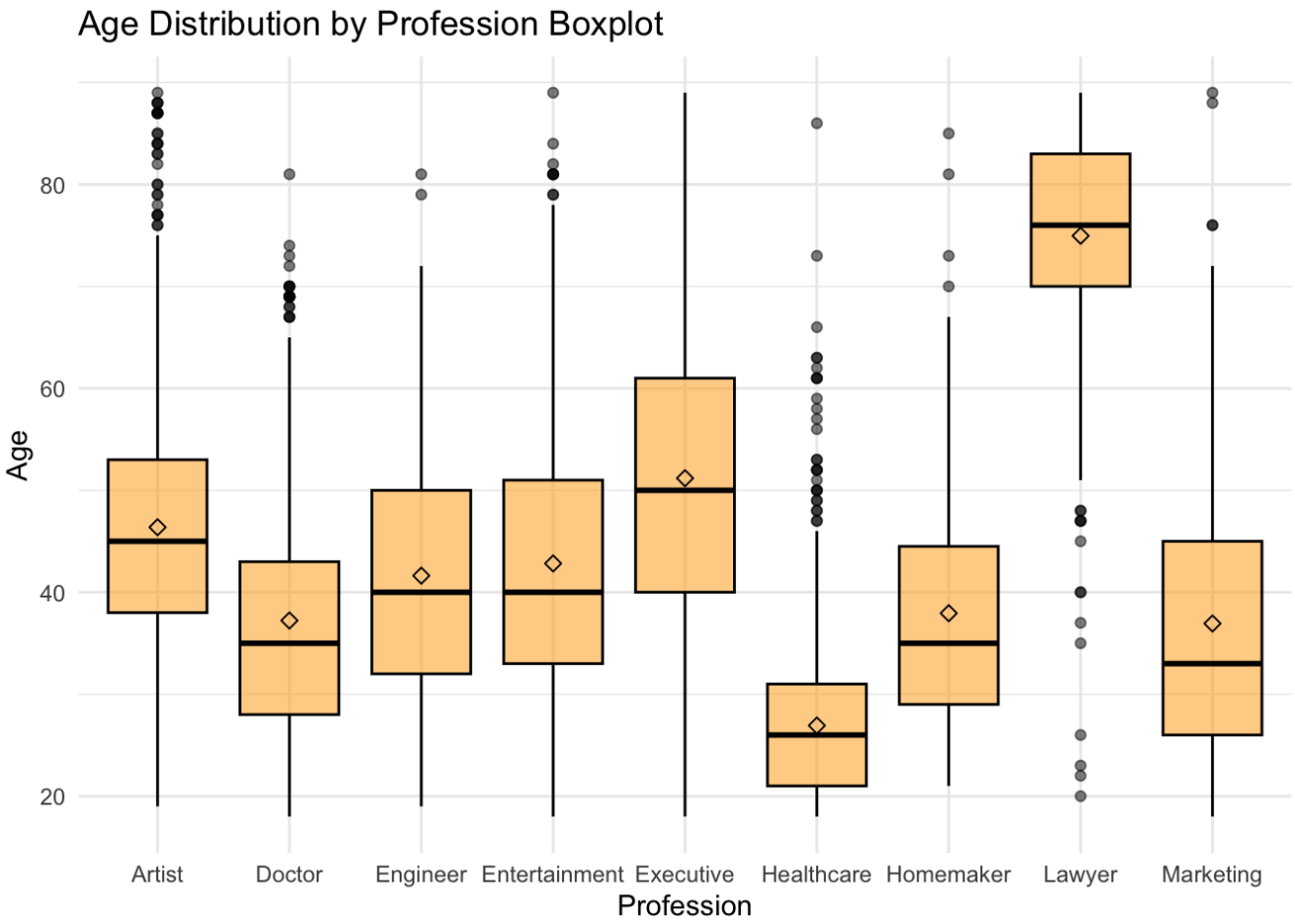
```
age_distrubition <- ggplot(clean_data, aes(x=Age)) +
  geom_histogram(color="black", fill="cadetblue3", bins = 30) +
  ggtitle("Distribution of Customer Age Histogram")
age_distrubition
```



Age Distribution by Profession Boxplot

It is observed from the boxplot that the majority of healthcare, medical (doctor), and marketing professionals in the dataset are relatively younger (below 40 years old), whereas law professionals tend to be relatively older (above 60 years old).

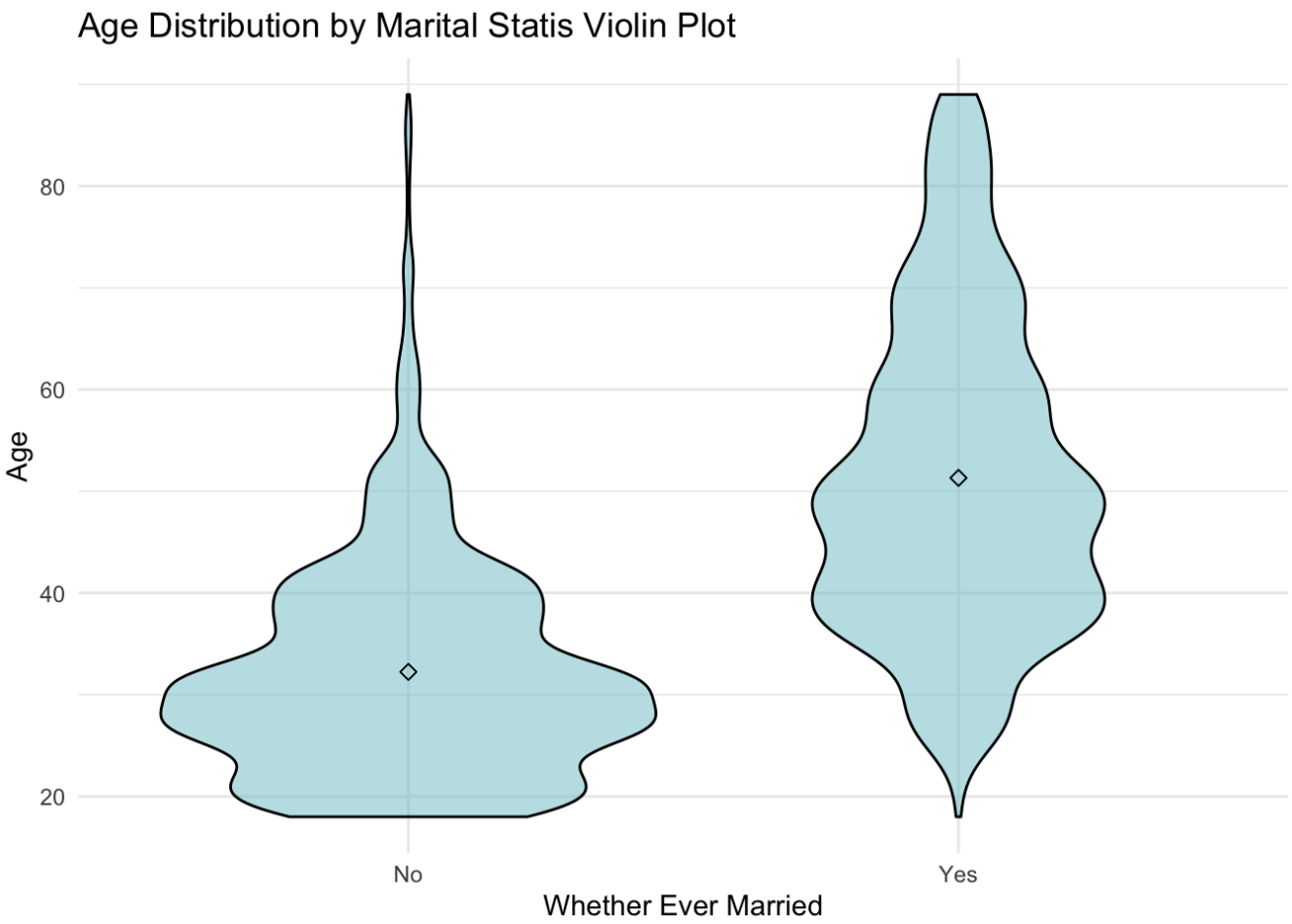
```
ggplot(clean_data, aes(x = Profession, y= Age)) +
  geom_boxplot(fill = "orange", color = "black", alpha = 0.5) +
  labs(title = "Age Distribution by Profession Boxplot", x = "Profession", y = "Age") +
  stat_summary(fun.y=mean, geom="point", shape=23, size=2)
```



Age Distribution by Marital Status Violin Plot

From the violin plot, it is evident that customers in the dataset who are not ever married tend to be younger than those who are ever married.

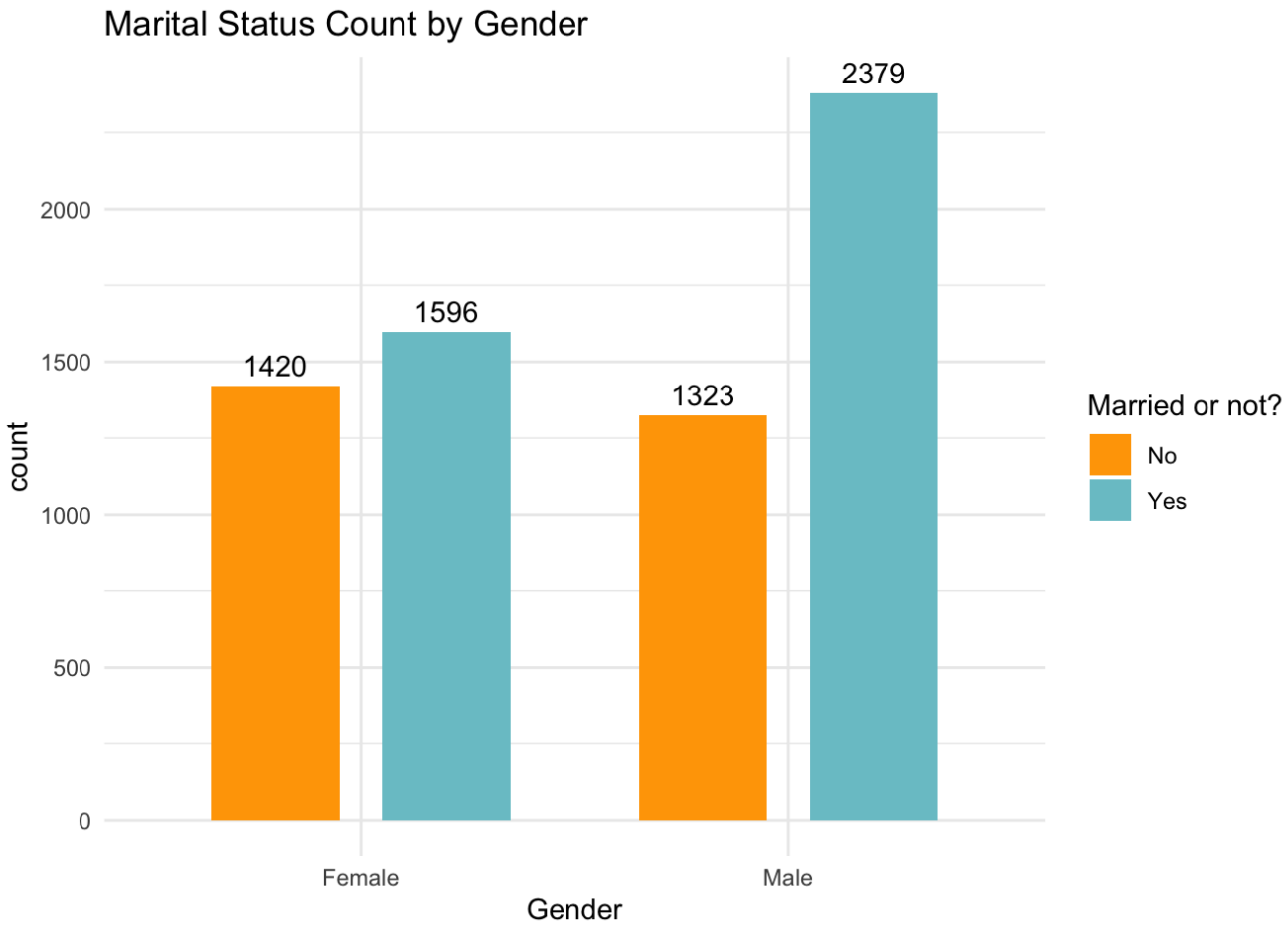
```
ggplot(clean_data, aes(x = Ever_Married, y= Age)) +
  geom_violin(fill = "cadetblue3", color = "black", alpha = 0.5) +
  labs(title = "Age Distribution by Marital Statis Violin Plot", x = "Whether Ever Marrie
d", y = "Age") +
  stat_summary(fun.y=mean, geom="point", shape=23, size=2)
```



Marital States by Gender Bar Chart

Based on the side-by-side bar chart, several observations can be made: firstly, there are more women who are not ever married than men. Secondly, there are more ever married men than women. Additionally, it is evident that the dataset contains more male customers than female customers.

```
gender_marriage_count <- clean_data |>
  group_by(Gender, Ever_Married) |>
  summarize(count = n(), .groups = "drop")

side_by_side_bar_chart <- ggplot(gender_marriage_count, aes(x = Gender, y = count, fill =
Ever_Married)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.6) +
  labs(title = "Marital Status Count by Gender", fill = "Married or not?") +
  geom_text(aes(label = count), position = position_dodge(width = 0.8), vjust = -0.5)+
  scale_fill_manual(values = c("Yes" = "cadetblue3", "No" = "orange"))

side_by_side_bar_chart
```



Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org (https://ggplot2.tidyverse.org).

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686).