Assignment 1

2025-10-07

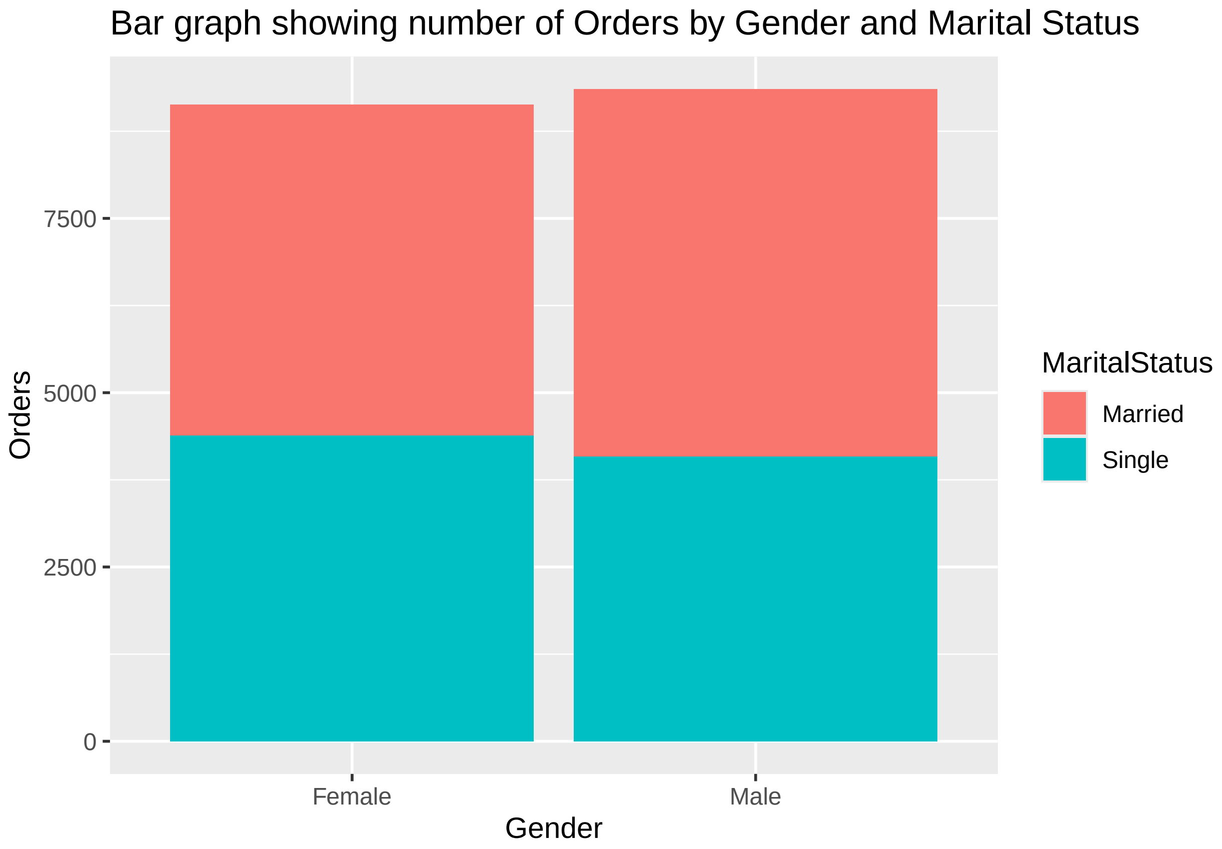
Importing data and merging required sheets.

```r
path <- 'C:\\Users\\Northwest-mwea\\Downloads\\BA Dataset.xlsx'

sheets <- c('Products', 'Customers', 'Orders')

# Merging vertically
sales <- lapply(sheets, function(s){
  read_excel(path, sheet = s) |>
    mutate(sheet = s, .before = 1)
}) |>
  bind_rows() |>
  mutate(
    Year = year(BirthDate),
    Age = 2025 - Year
  )

sales
```

```
> # A tibble: 79,488 x 47
>    sheet    ProductKey ProductAlternateKey ProductName        ProductSubcategory
>    <chr>         <dbl> <chr>               <chr>              <chr>
>  1 Products          1 AR-5381             Adjustable Race    <NA>
>  2 Products          2 BA-8327             Bearing Ball       <NA>
>  3 Products          3 BE-2349             BB Ball Bearing    <NA>
>  4 Products          4 BE-2908             Headset Ball Bear~ <NA>
>  5 Products          5 BL-2036             Blade              <NA>
>  6 Products          6 CA-5965             LL Crankarm        <NA>
>  7 Products          7 CA-6738             ML Crankarm        <NA>
>  8 Products          8 CA-7457             HL Crankarm        <NA>
>  9 Products          9 CB-2903             Chainring Bolts    <NA>
> 10 Products         10 CN-6137             Chainring Nut      <NA>
> # i 79,478 more rows
> # i 42 more variables: ProductCategoryName <chr>, StandardCost <dbl>,
> #   Color <chr>, ListPrice <dbl>, Size <chr>, SizeRange <chr>, Weight <dbl>,
> #   ProductLine <chr>, Class <chr>, Style <chr>, ModelName <chr>,
> #   Description <chr>, CustomerKey <dbl>, Title <chr>, FirstName <chr>,
> #   LastName <chr>, BirthDate <dttm>, MaritalStatus <chr>, Gender <chr>,
> #   EmailAddress <chr>, YearlyIncome <dbl>, TotalChildren <dbl>, ...
```

Stacked Bar chart of number of orders by Marital Status and Gender

```
sales |>
  select(MaritalStatus, Gender) |>
  drop_na() |>
  ggplot(aes(x = Gender, fill = MaritalStatus)) +
  geom_bar() +
  labs(
    title = 'Bar graph showing number of Orders by Gender and Marital Status',
    y = 'Orders'
  )
```



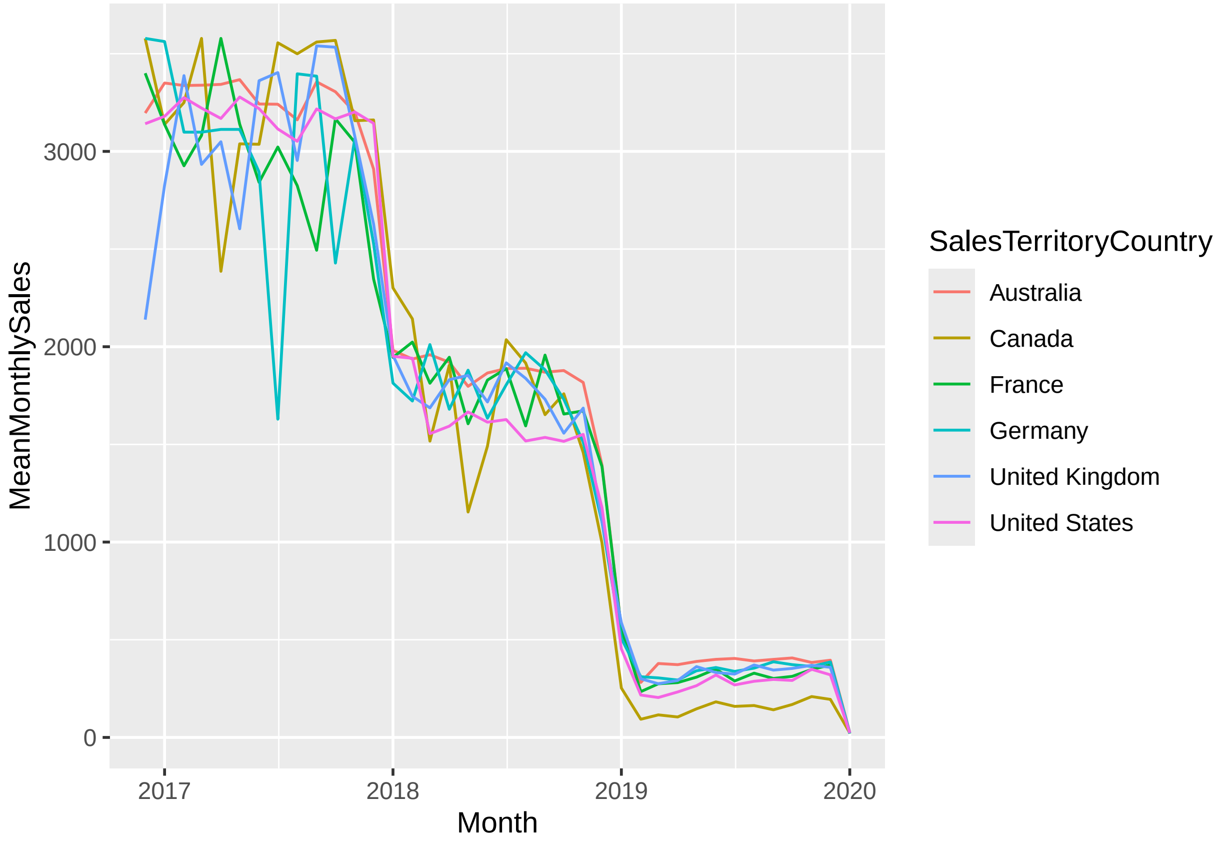
Based on the total length of each bar, we can conclude that the highest number of orders were received from males.

Comparing color segments across gender based on the absolute length of the segments at the baseline, we can conclude that single females had more number of orders than single males.

Also married males had more number of orders compared to married females because the married male's segment baseline was below that of married females and it also close above the segment of married females.

Grouped line chart of sales amount by order month for each SalesTerritoryCountry

```r
sales |>
  mutate(Date = as.Date(OrderDate), Month = floor_date(OrderDate, 'month')) |>
  group_by(SalesTerritoryCountry, Month) |>
  summarise(MeanMonthlySales = mean(SalesAmount, na.rm = TRUE), .groups = 'drop_last') |>
  ungroup() |>
  drop_na() |>
  ggplot(aes(x = Month, y = MeanMonthlySales, colour = SalesTerritoryCountry,
             group = SalesTerritoryCountry)) +
  geom_line()
```



The line graph shows that the mean monthly sales, at the beginning of year 2017, for the six territory countries was at climax despite United Kingdom having the lowest sales (Below 3000).

There was no clear direction on the trend of sales for every territory country between year 2017 and 2018 except for Germany whose sales spiked lower than others mid 2017 (Below 2000). There was a huge drop on monthly sales for every country across the years, with Canada recording the highest drop on sales, (Below 1500), between late 2017 and mid 2018.

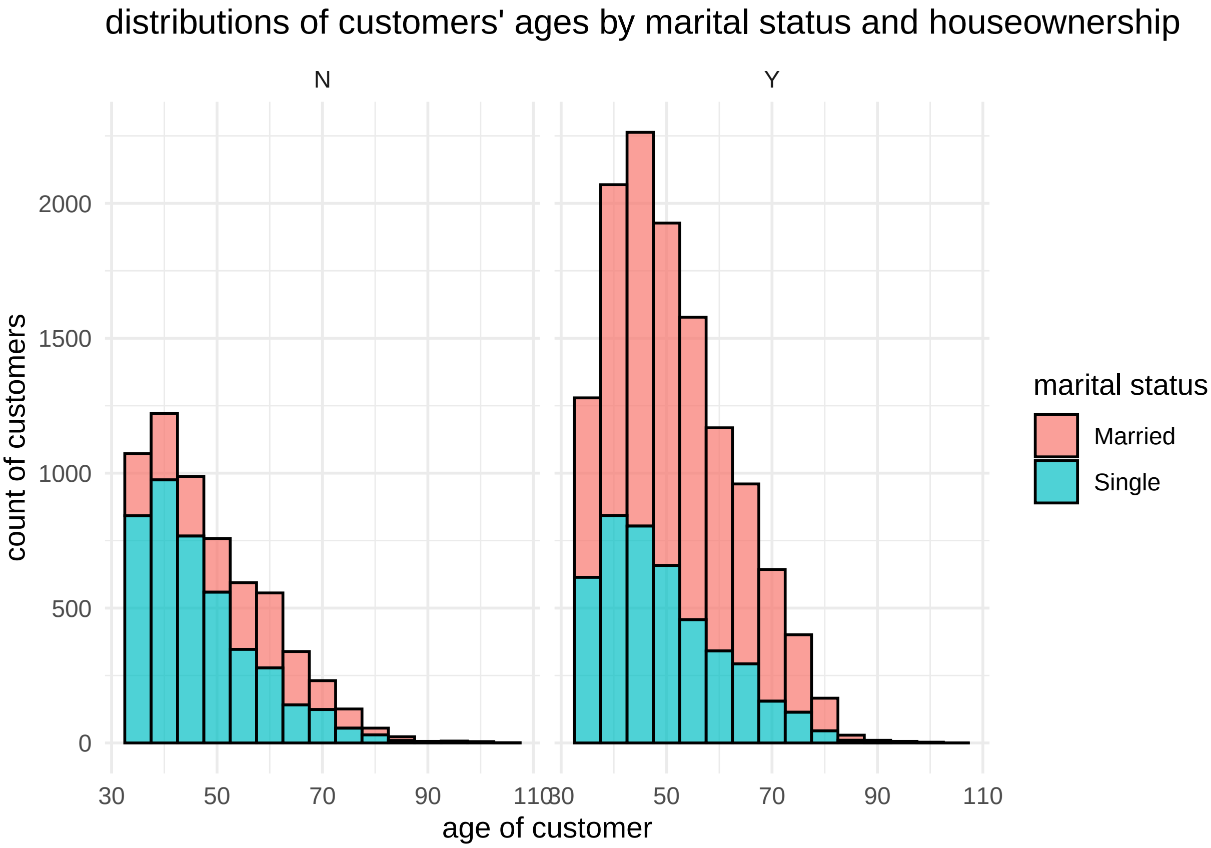
The second phase of monthly sales drop for all the territory countries took place between late 2018 and the beginning of year 2019.

There was no significant growth on sales as the chart was moving horizontally(constant) between 2019 and 2020.

We can conclude that there was a declining mean monthly sales in all the territory countries between year 2017 to 2020.

Histogram of customers' ages color bars by MaritalStatus and facet wrap by HouseOwnershipFlag

```r
sales |>
  select(Age, MaritalStatus, HouseOwnerFlag) |>
  drop_na() |>
  ggplot(aes(x = Age, fill = MaritalStatus)) +
    geom_histogram(binwidth = 5,color = "black",alpha = 0.7) +
    facet_wrap(~ HouseOwnerFlag) +
    labs(title="distributions of customers' ages by marital status and houseownership",
         x="age of customer",
         y="count of customers",
         fill="marital status") +
    theme_minimal()
```



The panel on the left shows customers who do not own a house.

The majority of these customers are younger, with a peak count in the 30-40 age range and a steady drop from ages 50 to 90.

Within this group, single customers are more prevalent than married customers, especially in the younger age brackets.

The panel on the right shows customers who own a house.

The age distribution for this group is centered on a higher age range, with a peak count between 50 and 60 years old.

In this group, married customers significantly outnumber single customers across all age ranges.

The data suggests a strong positive correlation between age, marital status, and houseownership.

Younger, single customers are more likely to not own a house, while older, married customers are more likely to be homeowners.

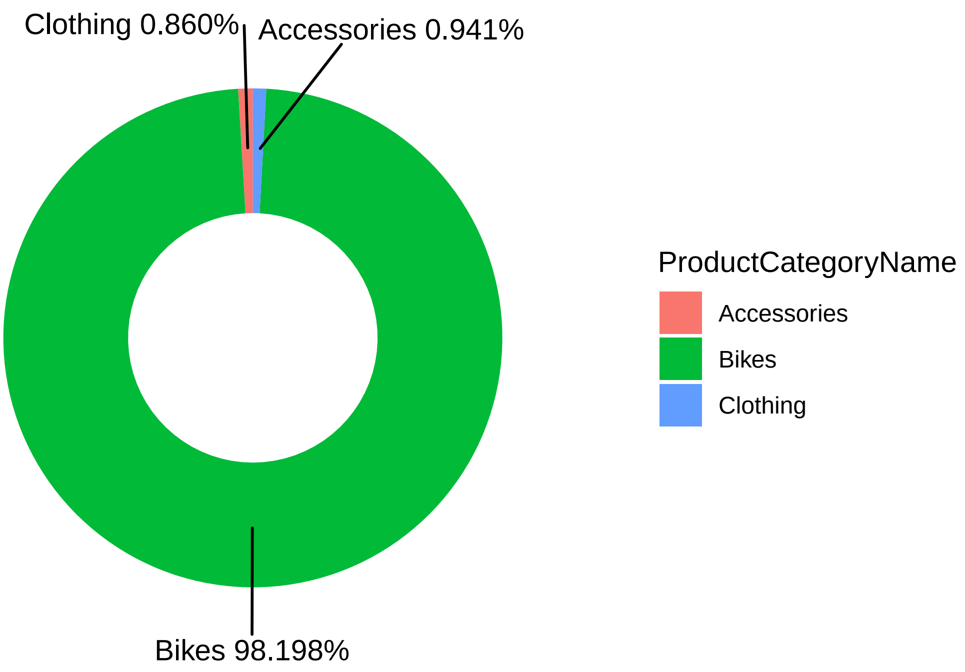
Donut chart of ProductCategoryName vs SalesAmount

Importing and merging dataset in required format

```r
# Performing left join merge style
products <- read_excel(path, sheet = "Products")
orders <- read_excel(path, sheet = "Orders")

Category_Sales <- orders |>
  left_join(products, by = c("ProductStandardCost" = "StandardCost")) |>
  select(ProductStandardCost, ProductCategoryName, SalesAmount)

Category_Sales
```

```
> # A tibble: 174,882 x 3
>    ProductStandardCost ProductCategoryName SalesAmount
>                  <dbl> <chr>                     <dbl>
>  1               2171. Bikes                     3578.
>  2               2171. Bikes                     3578.
>  3               2171. Bikes                     3578.
>  4               2171. Bikes                     3578.
>  5               2171. Bikes                     3578.
>  6               1912. Bikes                     3400.
>  7               1912. Bikes                     3400.
>  8               1912. Bikes                     3400.
>  9               1912. Bikes                     3400.
> 10               1912. Bikes                     3400.
> # i 174,872 more rows
```

```r
Category_Sales |>
  group_by(ProductCategoryName) |>
  summarise(Sales = sum(SalesAmount)) |>
  mutate(
    Fraction = Sales / sum(Sales),
    Label = paste0(ProductCategoryName, " ", scales::percent(Fraction))
  ) |>
  ggplot(aes(x = 2, y = Fraction, fill = ProductCategoryName)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  theme_void() +
  geom_text_repel(aes(label = Label, y = cumsum(Fraction) - Fraction / 2),
                  nudge_x = 1, show.legend = FALSE, size = 4) +
  xlim(c(0.5, 3)) +
  guides(fill = guide_legend(title = "ProductCategoryName")) +
  labs(title = 'Donut chart of the sale of different product categoires',
       caption = 'There were no sales of the components category')
```



Donut chart of the sale of different product categoires

There were no sales of the components category

The chart shows that Bikes make up the vast majority of sales, accounting for 98.198% of the total.

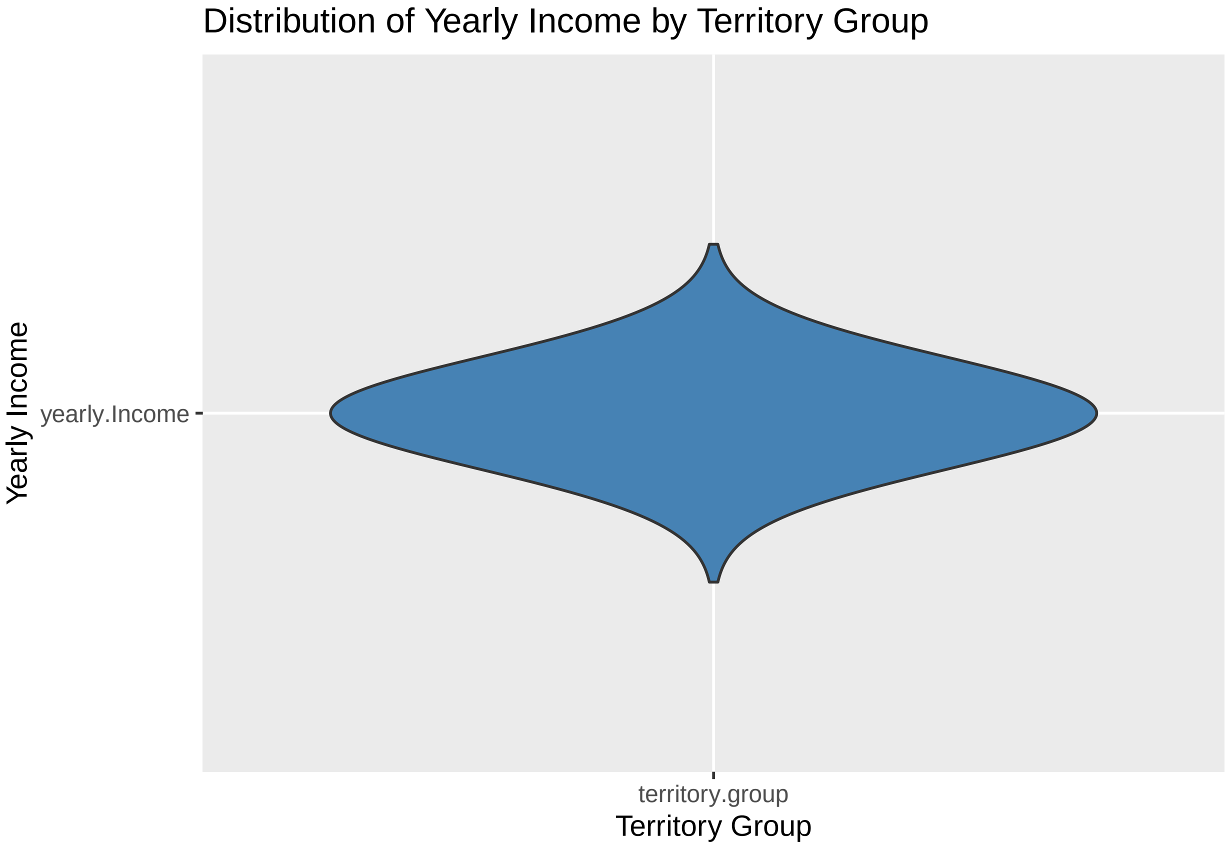
The other categories, accessories-0.941% and clothing-0.860%, represent a much smaller fraction of sales.

We can therefore conclude that Bikes are the best selling product.

The caption states that there was an additional Product Category called components. These had no sales at all and were therefore excluded from the graph.

Violin plot of YearlyIncome vs SalesTerritoryGroup

```
ggplot(sales, aes(x = 'territory.group', y = 'yearly.Income', fill = 'territory.group')) +
  geom_violin(trim =FALSE, fill='steelblue')+
  labs(
    title = "Distribution of Yearly Income by Territory Group",
    x = "Territory Group",
    y = "Yearly Income"
  ) +
  scale_fill_brewer()
```



Heatmap of Total sales by SalesTerritoryCountry and EnglishEducation

Importing and merging data set in required format

```
Customers <- read_excel(path, sheet = 'Customers')

Country_Education_sales <- inner_join(Customers, orders, by = 'CustomerKey') |>
  select(SalesTerritoryCountry, EnglishEducation, SalesAmount)

Country_Education_sales
```

```
> # A tibble: 60,398 x 3
>    SalesTerritoryCountry EnglishEducation SalesAmount
>    <chr>                 <chr>                  <dbl>
>  1 Australia             Bachelors            3400.
>  2 Australia             Bachelors            2320.
>  3 Australia             Bachelors              22.0
>  4 Australia             Bachelors            2384.
>  5 Australia             Bachelors              29.0
>  6 Australia             Bachelors               4.99
>  7 Australia             Bachelors              35.0
>  8 Australia             Bachelors              54.0
>  9 Australia             Bachelors            3375.
> 10 Australia             Bachelors            2320.
> # i 60,388 more rows
```

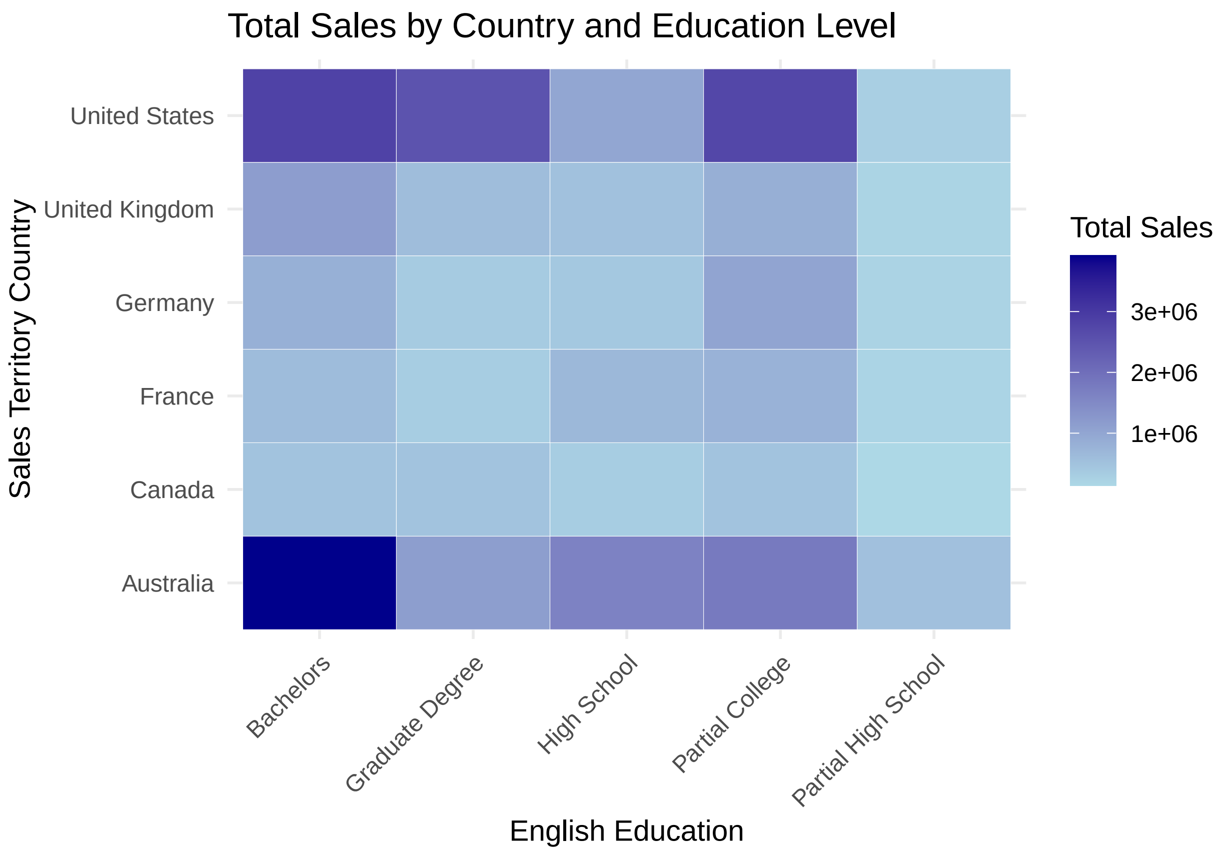
```
# Grouping Data
heatmap_data <- Country_Education_sales |>
  group_by(SalesTerritoryCountry, EnglishEducation) |>
  summarise(TotalSales = sum(SalesAmount, na.rm = TRUE), .groups = 'drop_last')

heatmap_data
```

```
> # A tibble: 30 x 3
> # Groups:   SalesTerritoryCountry [6]
>    SalesTerritoryCountry EnglishEducation    TotalSales
>    <chr>                 <chr>                    <dbl>
>  1 Australia             Bachelors             3922229.
>  2 Australia             Graduate Degree       1147062.
>  3 Australia             High School           1643721.
>  4 Australia             Partial College       1797196.
>  5 Australia             Partial High School    550791.
>  6 Canada                Bachelors              503163.
>  7 Canada                Graduate Degree        499428.
>  8 Canada                High School            329951.
>  9 Canada                Partial College        499950.
> 10 Canada                Partial High School    145354.
> # i 20 more rows
```

```
# Reshaping into matrix format necessary for heat map
heatmap_data |>
  pivot_wider(
    names_from = EnglishEducation,
    values_from = TotalSales
  )
```

```
> # A tibble: 6 x 6
> # Groups:   SalesTerritoryCountry [6]
>   SalesTerritoryCountry Bachelors `Graduate Degree` `High School`
>    <chr>                    <dbl>             <dbl>         <dbl>
> 1 Australia              3922229.          1147062.      1643721.
> 2 Canada                  503163.           499428.       329951.
> 3 France                  629162.           330788.       687521.
> 4 Germany                 840331.           362447.       424100.
> 5 United Kingdom         1167132.           603455.       544942.
> 6 United States          2838125.          2517379.      1007791.
> # i 2 more variables: `Partial College` <dbl>, `Partial High School` <dbl>
```

```r
# Creating Heatmap
heatmap_data |>
  ggplot(aes(x = EnglishEducation, y = SalesTerritoryCountry, fill = TotalSales)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Total Sales by Country and Education Level",
       x = "English Education",
       y = "Sales Territory Country",
       fill = "Total Sales") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Total Sales by Country and Education Level
Sales Territory Country
United States
United Kingdom
Germany
France
Canada
Australia
English Education
Bachelors
Graduate Degree
High School
Partial College
Partial High School
Total Sales
3e+06
2e+06
1e+06