day1_eda

```
churn <- read_csv(here('data/raw', 'Churn_Modelling.csv'))
head(churn)
```

```
# A tibble: 6 × 14
  RowNumber CustomerId Surname CreditScore Geography Gender
Age Tenure Balance
      <dbl>      <dbl> <chr>         <dbl> <chr>     <chr>
<dbl>  <dbl>   <dbl>
1         1   15634602 Hargra…         619 France    Female
42      2      0
2         2   15647311 Hill            608 Spain     Female
41      1  83808.
3         3   15619304 Onio            502 France    Female
42      8 159661.
4         4   15701354 Boni            699 France    Female
39      1      0
5         5   15737888 Mitche…         850 Spain     Female
43      2 125511.
6         6   15574012 Chu             645 Spain     Male
44      8 113756.
# ℹ 5 more variables: NumOfProducts <dbl>, HasCrCard <dbl>,
#   IsActiveMember <dbl>, EstimatedSalary <dbl>, Exited <dbl>
```

```
nrow(churn)
```

```
[1] 10000
```

```
ncol(churn)
```

```
[1] 14
```

```
glimpse(churn)
```

```
Rows: 10,000
Columns: 14
$ RowNumber       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,
13, 14, 15, 16,…
$ CustomerId      <dbl> 15634602, 15647311, 15619304,
15701354, 15737888, 1557…
$ Surname         <chr> "Hargrave", "Hill", "Onio", "Boni",
"Mitchell", "Chu",…
$ CreditScore     <dbl> 619, 608, 502, 699, 850, 645, 822,
```

```
376, 501, 684, 528,…
$ Geography       <chr> "France", "Spain", "France", "France",
"Spain", "Spain…
$ Gender          <chr> "Female", "Female", "Female",
"Female", "Female", "Mal…
$ Age             <dbl> 42, 41, 42, 39, 43, 44, 50, 29, 44,
27, 31, 24, 34, 25…
$ Tenure          <dbl> 2, 1, 8, 1, 2, 8, 7, 4, 4, 2, 6, 3,
10, 5, 7, 3, 1, 9,…
$ Balance         <dbl> 0.00, 83807.86, 159660.80, 0.00,
125510.82, 113755.78,…
$ NumOfProducts   <dbl> 1, 1, 3, 2, 1, 2, 2, 4, 2, 1, 2, 2, 2,
2, 2, 2, 1, 2, …
$ HasCrCard       <dbl> 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1,
0, 1, 0, 1, 1, …
$ IsActiveMember  <dbl> 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0,
0, 1, 1, 0, 1, …
$ EstimatedSalary <dbl> 101348.88, 112542.58, 113931.57,
93826.63, 79084.10, 1…
$ Exited          <dbl> 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 1, 0, …
```

```
colSums(is.na(churn))
```

```
      RowNumber      CustomerId         Surname
CreditScore       Geography
              0               0               0
0               0
         Gender             Age          Tenure
Balance   NumOfProducts
              0               0               0
0               0
      HasCrCard  IsActiveMember EstimatedSalary
Exited
              0               0               0
0
```

```
summary(churn)
```

```
   RowNumber       CustomerId         Surname
CreditScore
 Min.   :    1   Min.   :15565701   Length:10000       Min.
:350.0
 1st Qu.: 2501   1st Qu.:15628528   Class :character   1st
Qu.:584.0
 Median : 5000   Median :15690738   Mode  :character   Median
:652.0
```

```
 Mean   : 5000   Mean   :15690941                      Mean
:650.5
 3rd Qu.: 7500   3rd Qu.:15753234                      3rd
Qu.:718.0
 Max.   :10000   Max.   :15815690                      Max.
:850.0
  Geography            Gender               Age
Tenure
 Length:10000       Length:10000       Min.   :18.00   Min.
: 0.000
 Class :character   Class :character   1st Qu.:32.00   1st
Qu.: 3.000
 Mode  :character   Mode  :character   Median :37.00   Median
: 5.000
                                       Mean   :38.92   Mean
: 5.013
                                       3rd Qu.:44.00   3rd
Qu.: 7.000
                                       Max.   :92.00   Max.
:10.000
    Balance       NumOfProducts    HasCrCard
IsActiveMember
 Min.   :     0   Min.   :1.00   Min.   :0.0000   Min.
:0.0000
 1st Qu.:     0   1st Qu.:1.00   1st Qu.:0.0000   1st
Qu.:0.0000
 Median : 97199   Median :1.00   Median :1.0000   Median
:1.0000
 Mean   : 76486   Mean   :1.53   Mean   :0.7055   Mean
:0.5151
 3rd Qu.:127644   3rd Qu.:2.00   3rd Qu.:1.0000   3rd
Qu.:1.0000
 Max.   :250898   Max.   :4.00   Max.   :1.0000   Max.
:1.0000
 EstimatedSalary         Exited
 Min.   :    11.58   Min.   :0.0000
 1st Qu.: 51002.11   1st Qu.:0.0000
 Median :100193.91   Median :0.0000
 Mean   :100090.24   Mean   :0.2037
 3rd Qu.:149388.25   3rd Qu.:0.0000
 Max.   :199992.48   Max.   :1.0000
```

Churn Rate: 20.4%

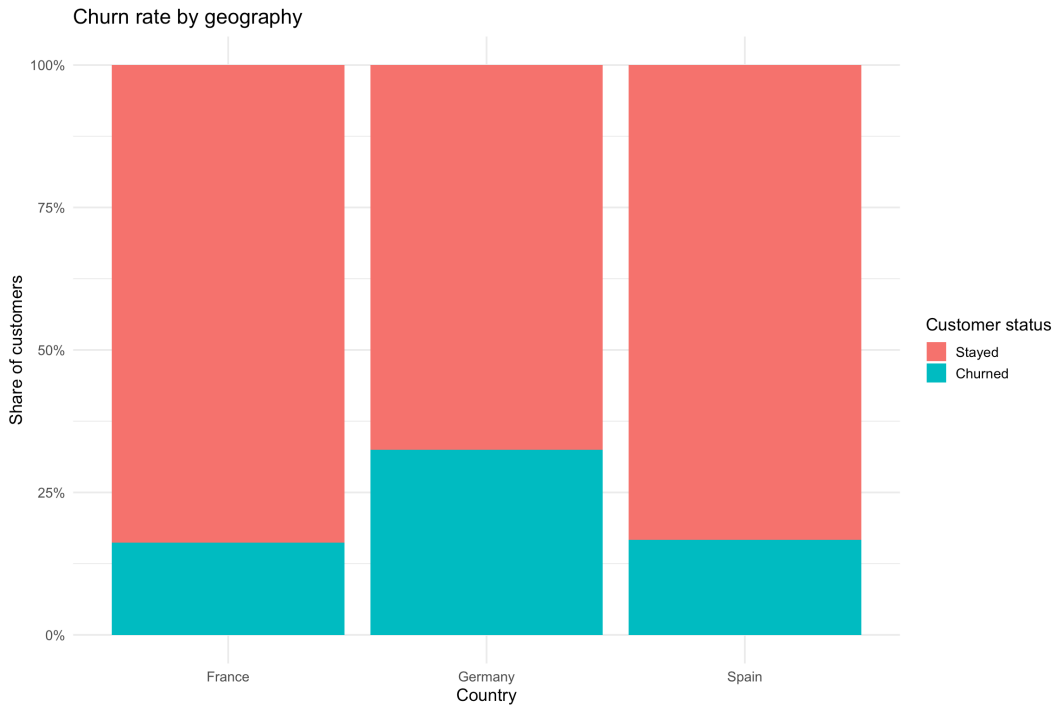
```
churn_rate <- churn|>
  count(Exited) |>
  mutate(pct = n/sum(n)*100)
```

```
churn_rate
```

```
# A tibble: 2 × 3
  Exited     n   pct
   <dbl> <int> <dbl>
1      0  7963  79.6
2      1  2037  20.4
```

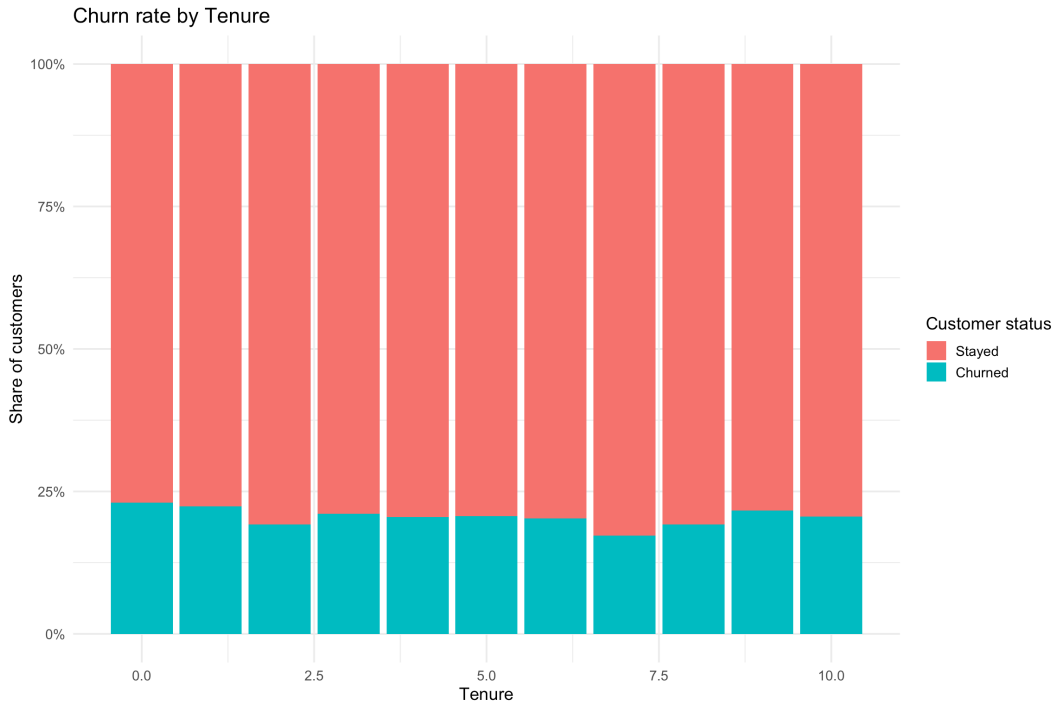
~80% of the people are happy and have stayed for this period.

```
ggplot(churn, aes(x = Geography, fill = factor(Exited))) +
  geom_bar(position = "fill") +
  scale_fill_discrete(
    name   = "Customer status",
    labels = c("0" = "Stayed", "1" = "Churned")
  ) +
  scale_y_continuous(labels = scales::label_percent(accuracy =
  labs(
    title = "Churn rate by geography",
    x = "Country",
    y = "Share of customers"
  ) +
  theme_minimal(base_size = 14)
```



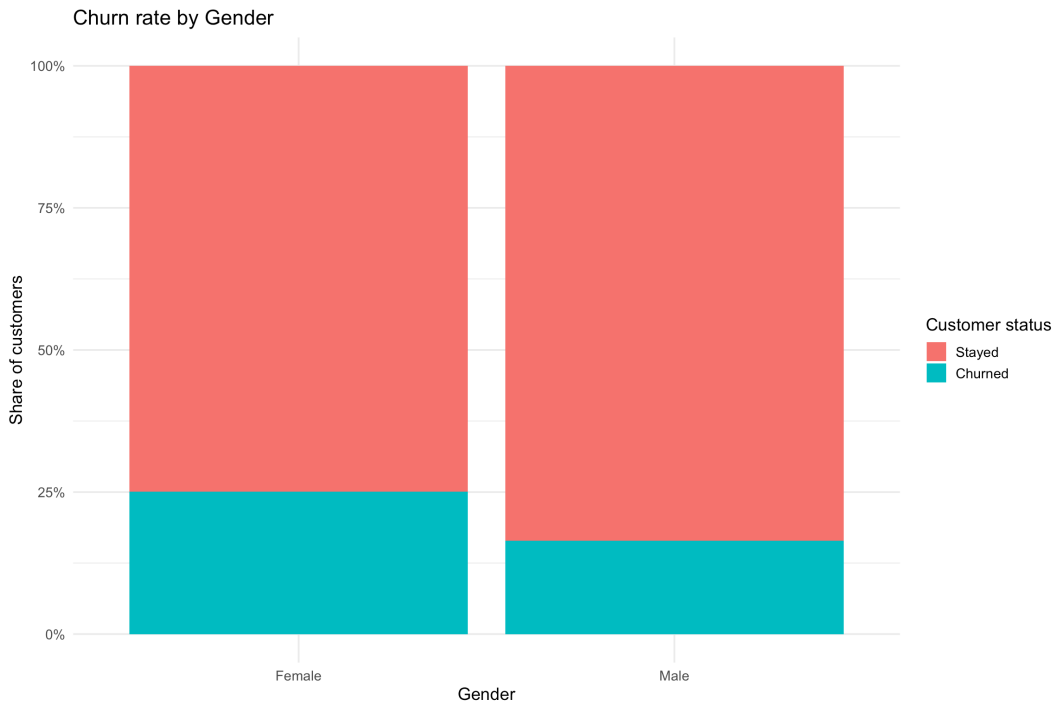
Germany has the maximum churned share of customers close to 30% , both France and Spain have closely equal share of customer who have churned (~15%).

```
ggplot(churn, aes(x = Tenure, fill = factor(Exited))) +
  geom_bar(position = "fill") +
  scale_fill_discrete(
    name   = "Customer status",
    labels = c("0" = "Stayed", "1" = "Churned")
  ) +
  scale_y_continuous(labels = scales::label_percent(accuracy =
  labs(
    title = "Churn rate by Tenure",
    x = "Tenure",
    y = "Share of customers"
  ) +
  theme_minimal(base_size = 14)
```



we see a uniform share of customers through all the tenure period, which mean customers remain loyal irrespective of whether being new or old(~18 to 22%).

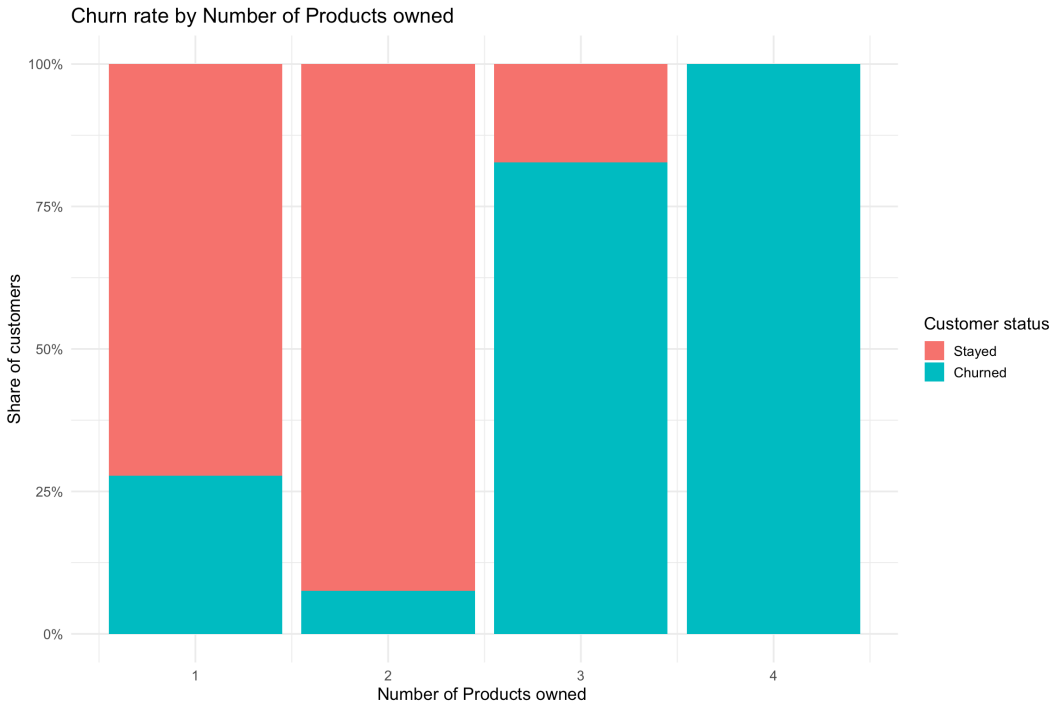
```
ggplot(churn, aes(x = Gender, fill = factor(Exited))) +
  geom_bar(position = "fill") +
  scale_fill_discrete(
    name   = "Customer status",
    labels = c("0" = "Stayed", "1" = "Churned")
  ) +
  scale_y_continuous(labels = scales::label_percent(accuracy =
  labs(
    title = "Churn rate by Gender",
    x = "Gender",
    y = "Share of customers"
  ) +
  theme_minimal(base_size = 14)
```



Female customer has churned more than male, with 25% for female and ~ 18% for male customers.

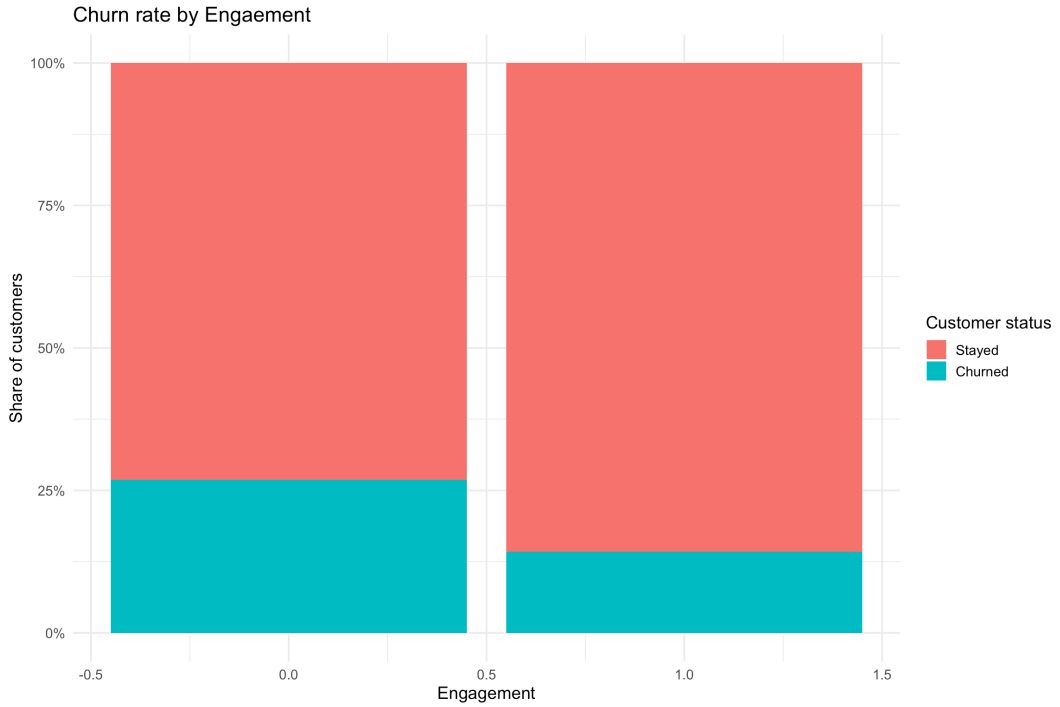
```
ggplot(churn, aes(x = NumOfProducts, fill = factor(Exited))) +
  geom_bar(position = "fill") +
  scale_fill_discrete(
    name   = "Customer status",
    labels = c("0" = "Stayed", "1" = "Churned")
  ) +
  scale_y_continuous(labels = scales::label_percent(accuracy =
```

```
  labs(
    title = "Churn rate by Number of Products owned",
    x = "Number of Products owned ",
    y = "Share of customers"
  ) +
  theme_minimal(base_size = 14)
```



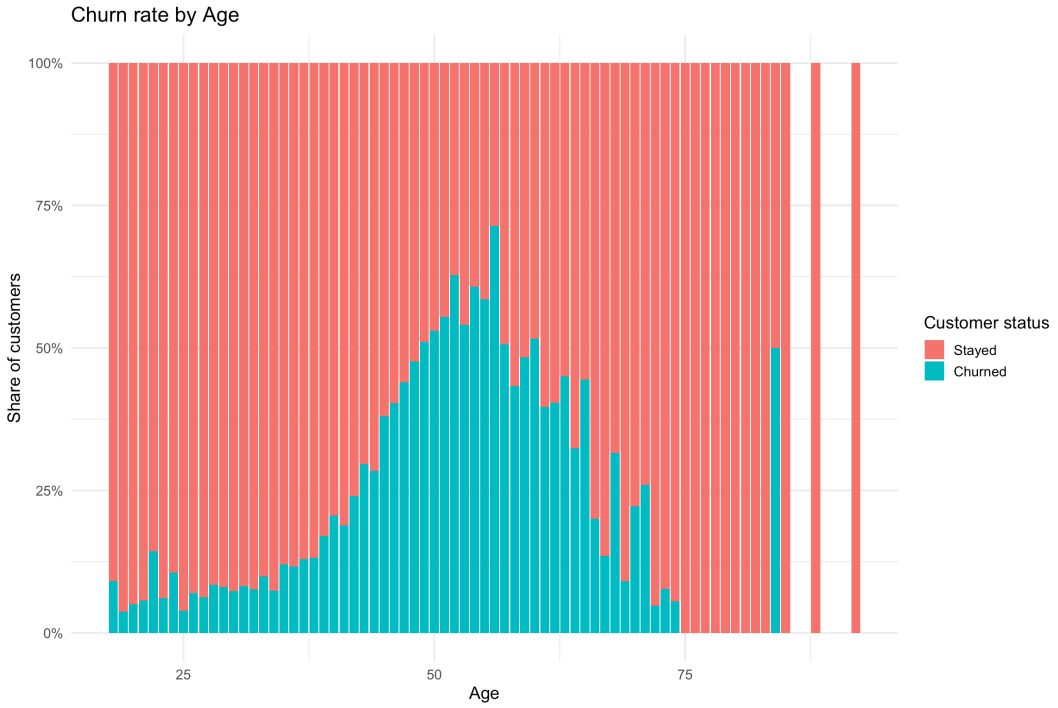
People with 3 and 4 products have churned most, with 4 product being 100 %. Its a huge driver for customer exiting.

```
ggplot(churn, aes(x = IsActiveMember, fill = factor(Exited))) +
  geom_bar(position = "fill") +
  scale_fill_discrete(
    name   = "Customer status",
    labels = c("0" = "Stayed", "1" = "Churned")
  ) +
  scale_y_continuous(labels = scales::label_percent(accuracy =
  labs(
    title = "Churn rate by Engaement",
    x = "Engagement",
    y = "Share of customers"
  ) +
  theme_minimal(base_size = 14)
```



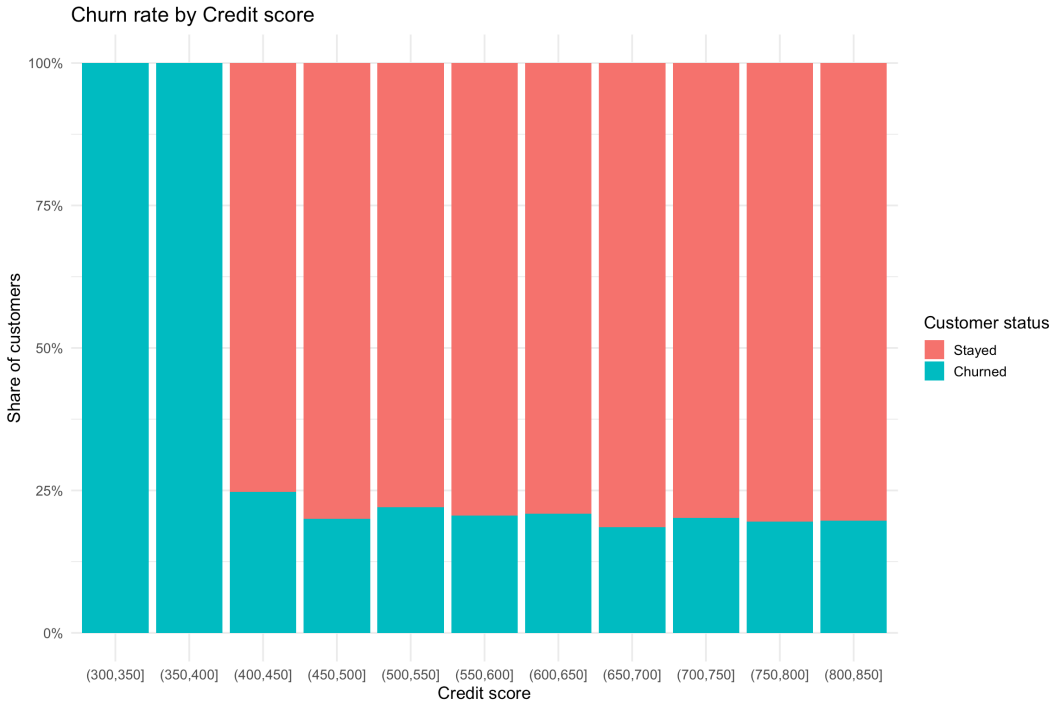
People who are not acitve have churned more than people who are acitve.

```
ggplot(churn, aes(x = Age, fill = factor(Exited))) +
  geom_bar(position = "fill") +
  scale_fill_discrete(
    name   = "Customer status",
    labels = c("0" = "Stayed", "1" = "Churned")
  ) +
  scale_y_continuous(labels = scales::label_percent(accuracy =
  labs(
    title = "Churn rate by Age",
    x = "Age",
    y = "Share of customers"
  ) +
  theme_minimal(base_size = 14)
```



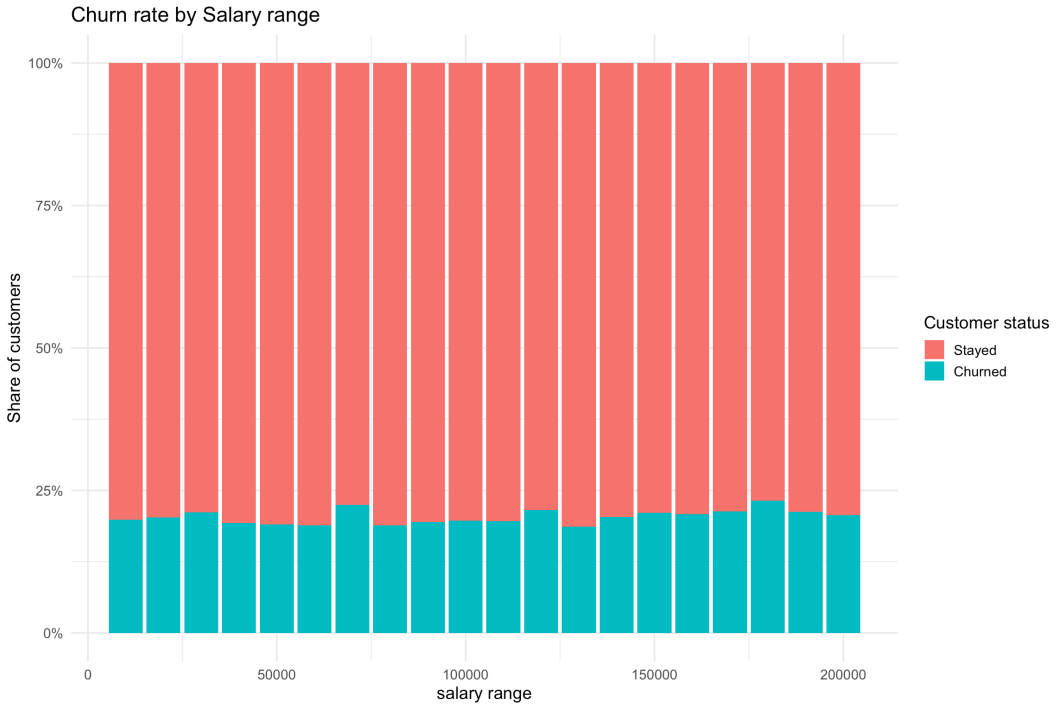
it seems to normal distributed, with mean close 50-55 yrs, which means people close to 50 to 55 yrs has churned most. it seems to have a outlier at ~80yrs, maybe due to people's death.

```
churn_df <- churn|>
  mutate(credit_bin = cut(CreditScore, breaks=seq(300,850,50)))

ggplot(churn_df, aes(x = credit_bin, fill = factor(Exited))) +
  geom_bar(position = "fill") +
  scale_fill_discrete(
    name   = "Customer status",
    labels = c("0" = "Stayed", "1" = "Churned")
  ) +
  scale_y_continuous(labels = scales::label_percent(accuracy =
  labs(
    title = "Churn rate by Credit score",
    x = "Credit score",
    y = "Share of customers"
  ) +
  theme_minimal(base_size = 14)
```



When the customers credit score is less than 400, their churning rate is very high close to 100% percent. the chruning rate decrease gradually as the credit score increases.

```
churn_df <- churn|>
  mutate(salary_bin = ceiling(EstimatedSalary/10000)*10000)

ggplot(churn_df, aes(x = salary_bin, fill = factor(Exited))) +
  geom_bar(position = "fill") +
  scale_fill_discrete(
    name   = "Customer status",
    labels = c("0" = "Stayed", "1" = "Churned")
  ) +
  scale_y_continuous(labels = scales::label_percent(accuracy =
  labs(
    title = "Churn rate by Salary range",
    x = "salary range",
    y = "Share of customers"
  ) +
  theme_minimal(base_size = 14)
```



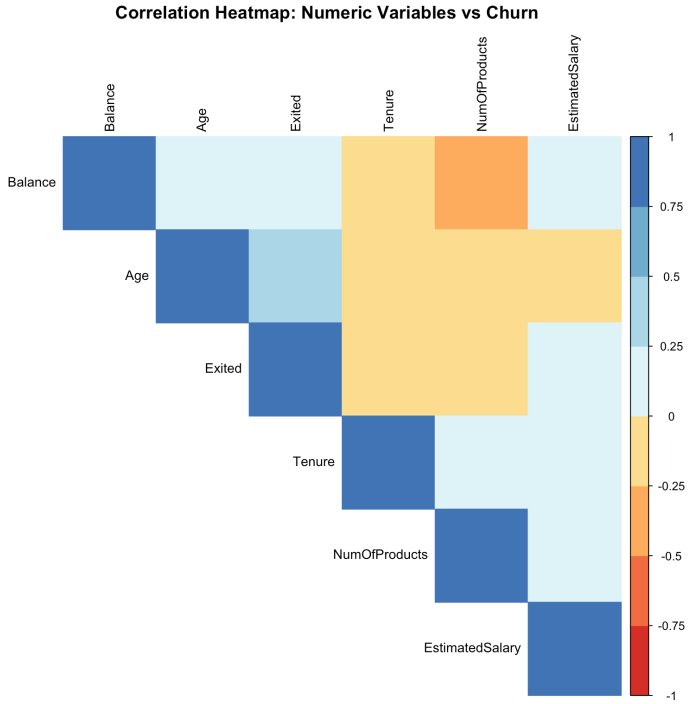
we see a uniform share of customers through all the Salary range, which mean customers remain loyal irrespective of their salary(~20 to 25%).

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
library(RColorBrewer)

# Select numeric vars for correlation
num_vars <- churn %>%
  select(Age, Tenure, Balance, NumOfProducts, EstimatedSalary,
  cor(use = "complete.obs")

# Heatmap
corrplot(num_vars,
         method = "color",
         type = "upper",
         order = "hclust",
         tl.cex = 0.9,
         tl.col = "black",
         col = brewer.pal(n = 8, name = "RdYlBu"),
```

```
        title = "Correlation Heatmap: Numeric Variables vs Chu
        mar = c(0,0,2,0))
```



“hypotheses for modeling” summary:

- Higher churn in Germany.
- Higher churn in females and mid-age customers.
- Very high churn when credit score < 400.
- Very high churn when people have more than 3 products.
- Higher churn in people who are not acitve.
- Tenure and salary look relatively flat vs churn.
- no serious multicollinearity problem among these numeric features, and only a weak-to-moderate relationship between the predictors and churn