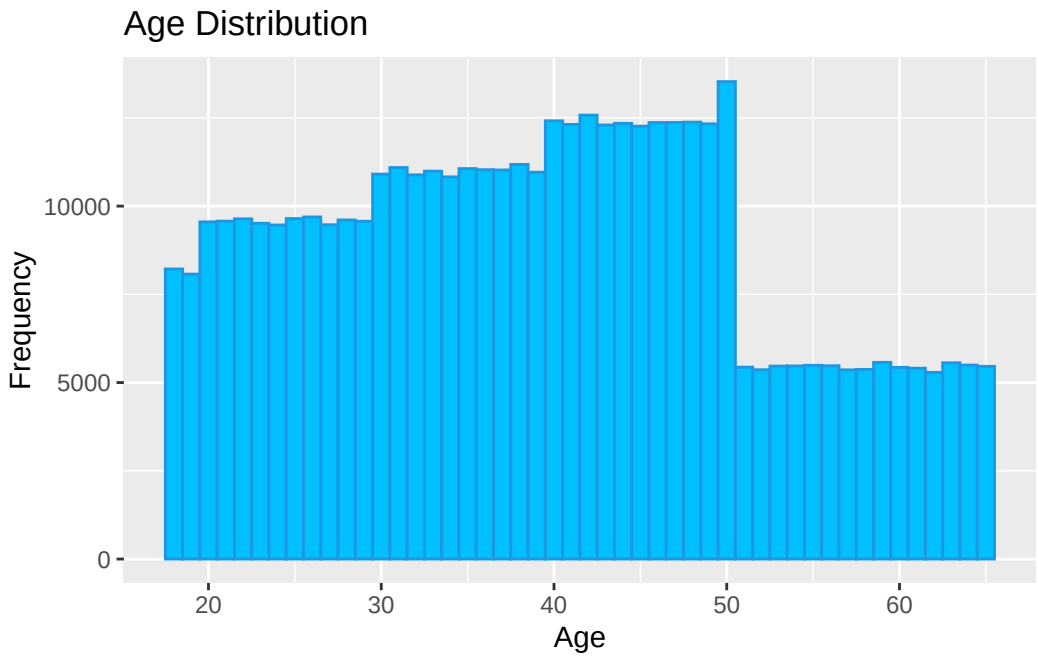
Exploratory Data Analysis

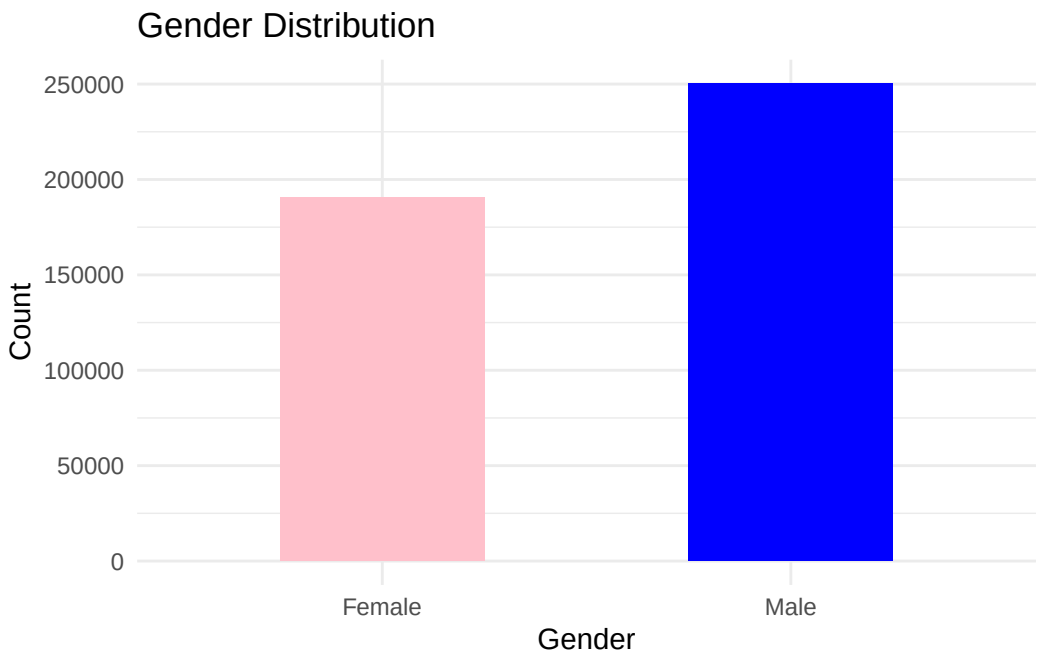
```
Attaching package: 'arrow'

The following object is masked from 'package:utils':

    timestamp


Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


Attaching package: 'sparklyr'

The following object is masked from 'package:stats':

    filter

# A tibble: 1 x 12
  CustomerID   Age Gender Tenure `Usage Frequency` `Support Calls`
       <dbl> <dbl>  <dbl>  <dbl>             <dbl>           <dbl>
1          1     1      0      1                 1               1
# i 6 more variables: `Payment Delay` <dbl>, `Subscription Type` <dbl>,
#   `Contract Length` <dbl>, `Total Spend` <dbl>, `Last Interaction` <dbl>,
#   Churn <dbl>
```



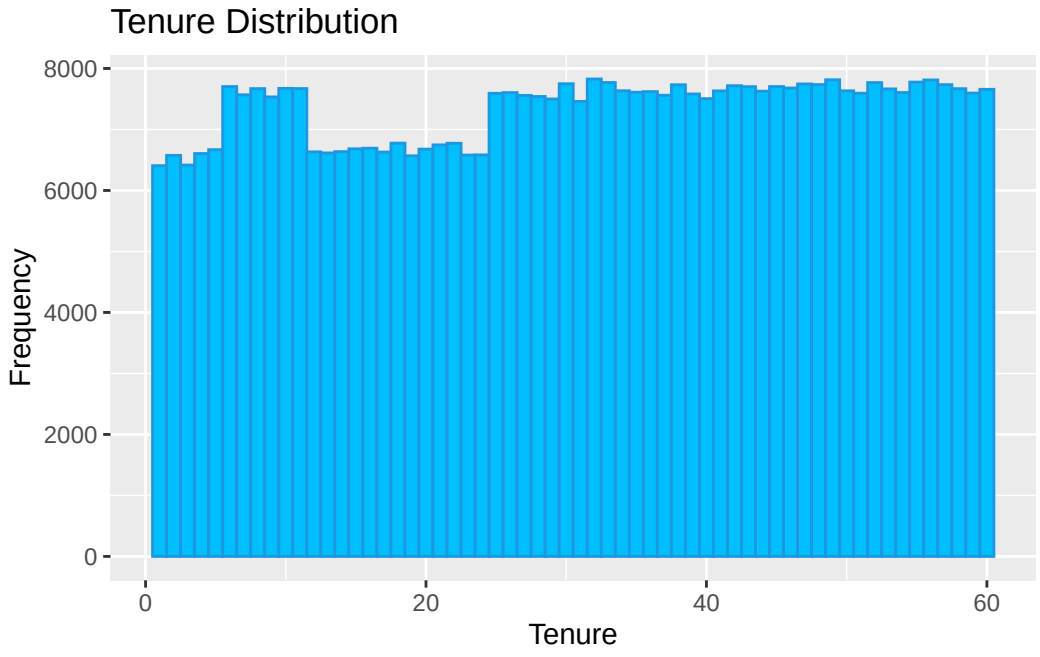

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   29.00   39.00   39.37   48.00   65.00
```

Looking at the distribution by age, a large portion of our dataset comes from aged 40-50 years old and a low portion comes from age 50-60 years old.

The statistic summary of age shows that the mean age is 39.37 and it's ranging from 18 to 65 years old.



From the Gender Distribution bar chart, we are able to see that we have more Male than Female in this dataset.

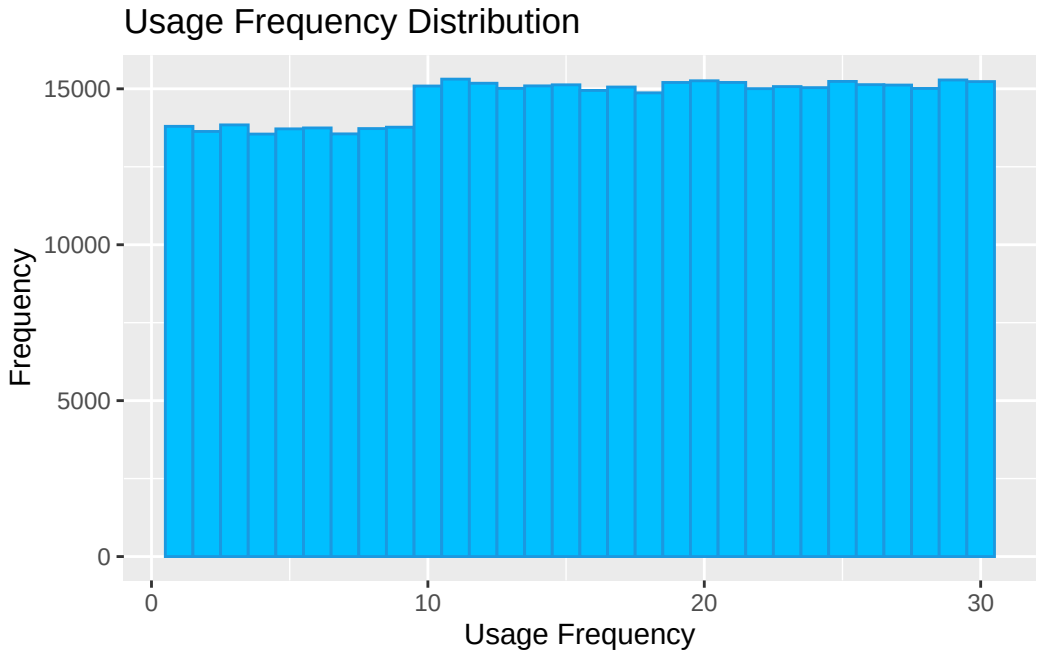



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
1.00   16.00   32.00   31.26   46.00   60.00
```

Looking at the Distribution by Tenure, it is apparent that the dataset exhibits a generally balanced distribution where majority of Tenure values hover around range of 7000 - 8000 in terms of count. However, for tenure values falling within the ranges of 10-20 and 0-5, there is a slight drop with counts ranging from 6000 to 7000, indicating a comparatively lower frequency of occurrences within these specific tenure ranges."

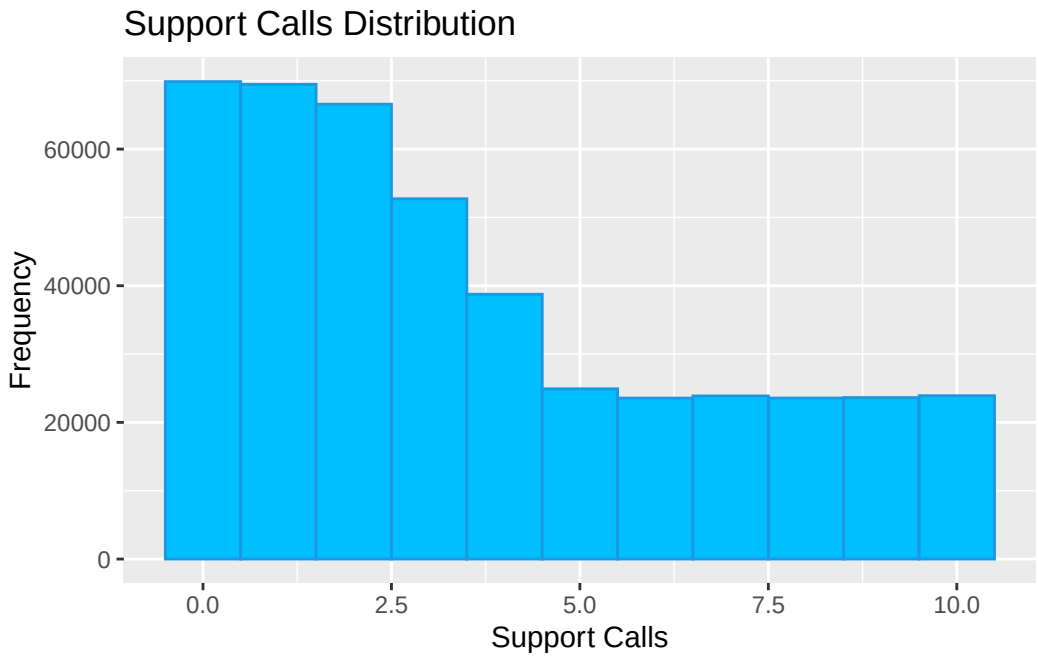
The statistic summary of tenure shows that the dataset mean tenure is 31.26 and it's ranging from 1 to 60 Tenure.




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    9.00   16.00   15.81   23.00   30.00
```

Looking at the Distribution by Usage Frequency, it is obvious that the dataset exhibits a generally balanced distribution where majority of Tenure values hover around range of 15000 in terms of count. However, for Usage Frequency values < 10, the counts is slightly lower.

The statistic summary of Usage Frequency shows that the dataset mean Usage Frequency is 15.81 and it's ranging from 1 to 30 Usage Frequency.




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   3.000   3.604   6.000  10.000
```

Looking at the Distribution by Support Calls, there is a downward trend in terms of counts as support calls increases. Support calls of 5 onward has a stable frequency of around 20000.

The statistic summary of Support Calls shows that the dataset mean Support Calls is 3.604 and it's ranging from 0 to 10 calls.