claims_analysis.R

rocka

2025-01-26

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
claims_df <- read.csv("claims.csv")

head(claims_df)
```

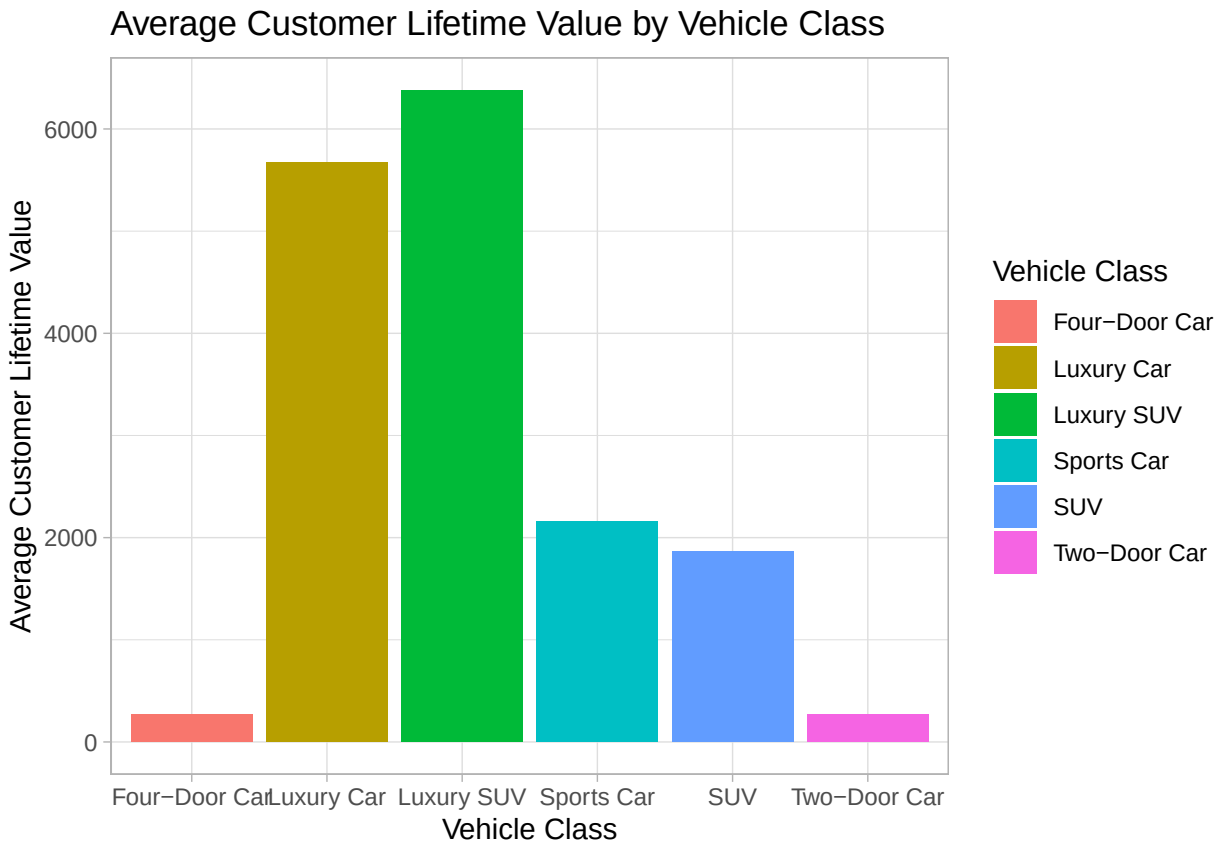
```
##   customer_id customer_state highest_education employment_status gender income
## 1     AA11235         Nevada          Bachelor     Medical Leave Female  11167
## 2     AA16582     Washington          Bachelor     Medical Leave   Male  14072
## 3     AA34092     California         Associate          Employed   Male  33635
## 4     AA56476        Arizona       High School          Employed Female  74454
## 5     AA69265         Nevada          Bachelor          Employed Female  60817
## 6     AA71604        Arizona            Master          Employed Female  87560
##   residence_type marital_status sales_channel coverage    policy vehicle_class
## 1       Suburban        Married        Branch    Basic  Personal  Two-Door Car
## 2       Suburban       Divorced         Agent    Basic  Personal Four-Door Car
## 3       Suburban        Married           Web Extended  Personal    Luxury SUV
## 4       Suburban         Single   Call Center    Basic Corporate Four-Door Car
## 5       Suburban         Single           Web  Premium  Personal Four-Door Car
## 6       Suburban        Married           Web Extended  Personal  Two-Door Car
##   vehicle_size monthly_premium months_policy_active months_since_last_claim
## 1      Midsize              73                   25                      0
## 2      Midsize              71                   27                     13
## 3      Midsize             240                   32                      1
## 4      Midsize              71                   39                     25
## 5      Midsize             103                   21                      3
## 6      Midsize              98                   17                      4
##   current_claim_amount total_claims total_claims_amount customer_lifetime_value
## 1                 1383            1                1383                    442
## 2                 1379            2                1992                    -75
## 3                 2633            2                3671                   4009
## 4                  906            2                1541                   1228
```

```
## 5                 1095            2                1760                    403
## 6                 1136            2                1828                   -162
```

```
#_____________________________________________________________________________________________
# Q1
# What impact does vehicle class have on the profitability of customers?
df1<-claims_df%>% group_by(vehicle_class)%>% summarise(n_customers=n(),
                                                       Average_CLV = mean(customer_lifetime_value),
                                                       Median_CLV = median(customer_lifetime_value),
                                                       pct_P=mean(customer_lifetime_value>0))

print(df1)
```

```
## # A tibble: 6 x 5
##   vehicle_class n_customers Average_CLV Median_CLV pct_P
##   <chr>               <int>       <dbl>      <dbl> <dbl>
## 1 Four-Door Car        3124        271.       176. 0.556
## 2 Luxury Car            119       5670.      5514  1
## 3 Luxury SUV            133       6382.      6142  0.992
## 4 SUV                  1246       1861.      1726. 0.846
## 5 Sports Car            335       2159.      1806  0.890
## 6 Two-Door Car         1292        269.       160. 0.539
```

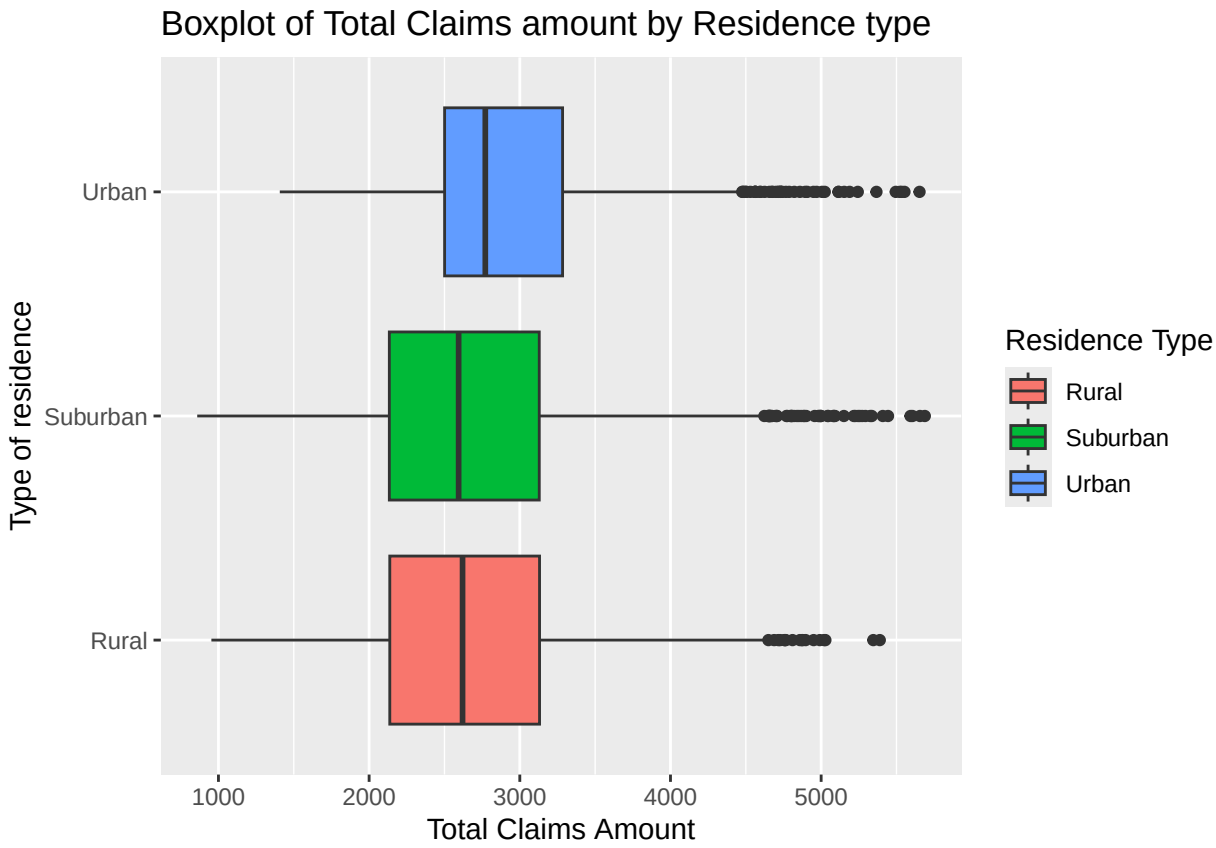
```
# This code adjusts the figure output size in the notebook
options(repr.plot.width=11, repr.plot.height=8)
ggplot(df1, aes(x = vehicle_class, y = Average_CLV, fill = vehicle_class)) +
  geom_col() +
  labs(title = "Average Customer Lifetime Value by Vehicle Class",
       x = "Vehicle Class",
       y = "Average Customer Lifetime Value",
       fill="Vehicle Class")+theme_light()
```



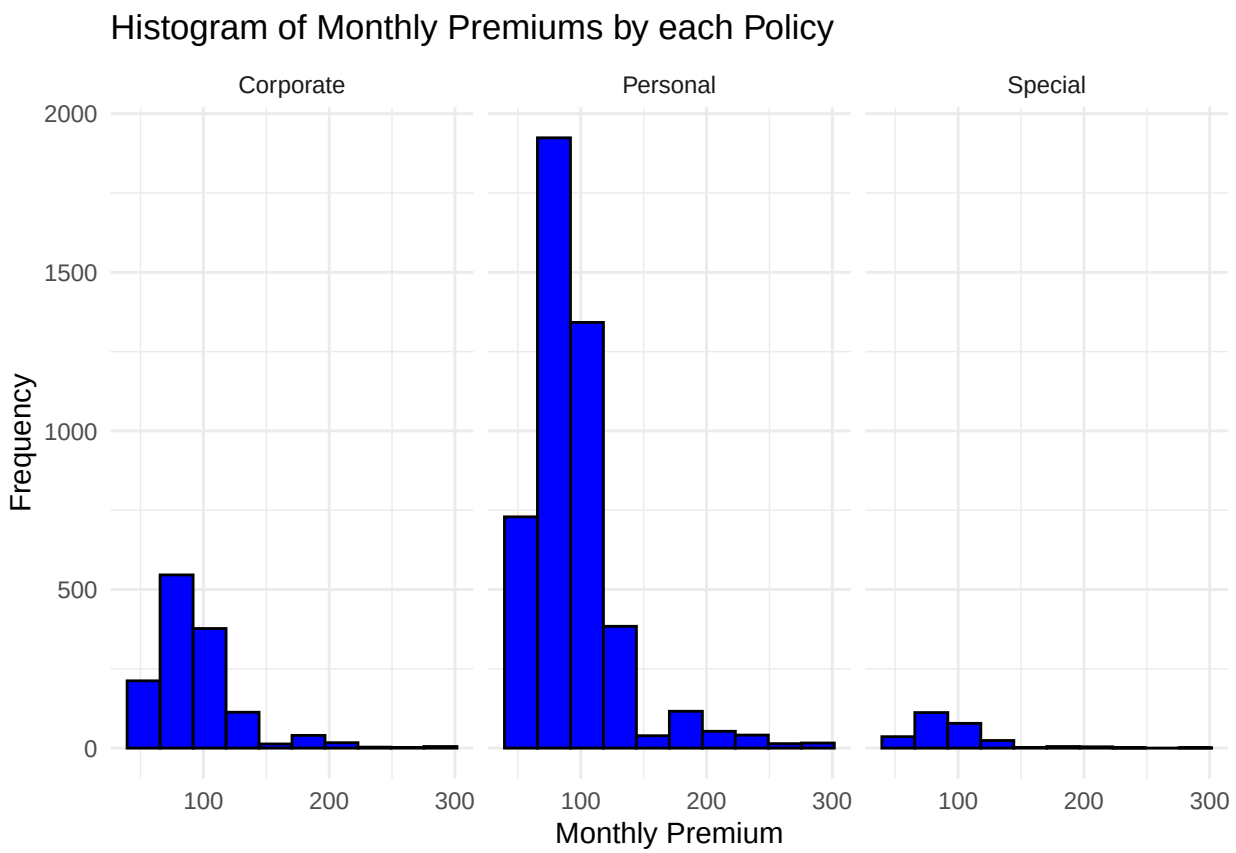

```
#-------------------------------------------------------------------------------------------------------

# Q2
# How does the total claim amount vary according to the type of residence?

(df2 <- claims_df %>%
    group_by(residence_type) %>%
    summarise(n_customers=n(),
              Profit_Ratio = mean(customer_lifetime_value>0),
              Avg_claims_amount=mean(total_claims_amount),
              Min_claims_amount=min(total_claims_amount),
              Max_claims_amount=max(total_claims_amount)))
```

```
## # A tibble: 3 x 6
##   residence_type n_customers Profit_Ratio Avg_claims_amount Min_claims_amount
##   <chr>                <int>        <dbl>             <dbl>             <int>
## 1 Rural                 1097        0.677             2691.               953
## 2 Suburban              3657        0.665             2688.               859
## 3 Urban                 1495        0.577             2924.              1407
## # i 1 more variable: Max_claims_amount <int>
```

```
options(repr.plot.width=11, repr.plot.height=8)
ggplot(claims_df,aes(x=residence_type,y=total_claims_amount,fill=residence_type))+
  geom_boxplot()+
  labs(title = "Boxplot of Total Claims amount by Residence type",
       x = "Type of residence",
       y = "Total Claims Amount",
       fill="Residence Type") +coord_flip()
```




```
#-------------------------------------------------------------------------------------------------

# Q3
# How does the distribution of monthly premiums differ among Corporate, Personal,
# and Special policyholders, and to what extent do customers across these policy types
# opt for premiums at or below the $100 mark?

ggplot(claims_df, aes(x = monthly_premium)) +
  geom_histogram(bins = 10, fill = "blue", color = "black") +
  facet_wrap(~policy)+
  labs(title = "Histogram of Monthly Premiums by each Policy ",
       x = "Monthly Premium",
       y = "Frequency") +
  theme_minimal()
```




```
(df3<- claims_df %>% group_by(policy) %>% summarise(n_customers=n(),
                                                    avg_monthly_premium = mean(monthly_premium),
                                                    max_monthly_premium = max(monthly_premium),
                                                    sd_monthly_premium = sd(monthly_premium),
                                                    pct_less_100 = mean(monthly_premium <= 100)))
```

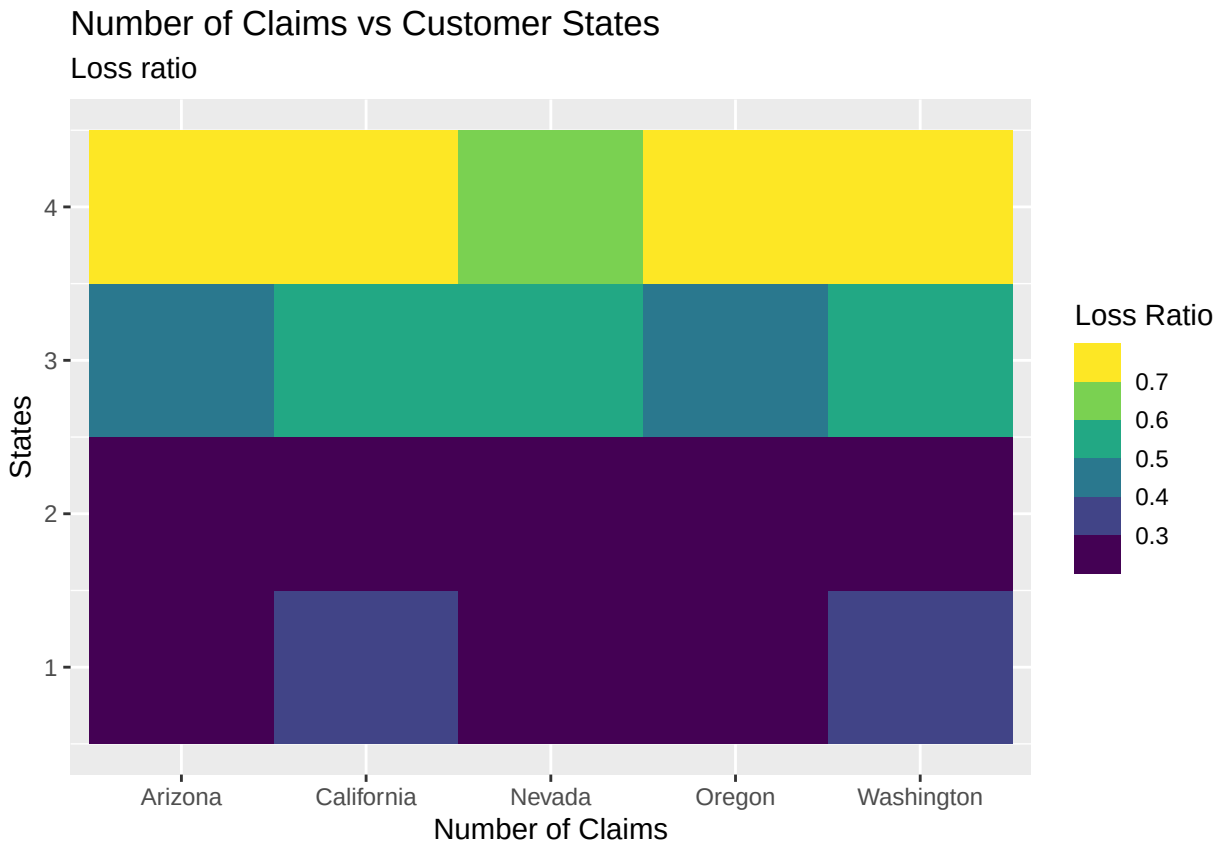
```
## # A tibble: 3 x 6
##   policy  n_customers avg_monthly_premium max_monthly_premium sd_monthly_premium
##   <chr>         <int>               <dbl>               <int>              <dbl>
## 1 Corpor~        1328                93.6                 285               34.4
## 2 Person~        4658                93.9                 297               35.7
## 3 Special         263                93.3                 283               33.0
## # i 1 more variable: pct_less_100 <dbl>
```

```
#_____________________________________________________________________________________________

# Q4
# Are there any geographical patterns in number of claims made and profitability?
(df4<- claims_df %>% group_by(customer_state,total_claims)%>%
    summarise(Avg_CLV_Loss = mean(customer_lifetime_value<0),
              n_customers = n(),
              Median_CLV = median(customer_lifetime_value),
              Average_CLV = mean(customer_lifetime_value))) %>%
  arrange(desc(Avg_CLV_Loss))
```

```
## `summarise()` has grouped output by 'customer_state'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 20 x 6
## # Groups:   customer_state [5]
##    customer_state total_claims Avg_CLV_Loss n_customers Median_CLV Average_CLV
##    <chr>                 <int>        <dbl>       <int>      <dbl>       <dbl>
##  1 Washington                4        0.759          29    -1382         -930.
##  2 Arizona                   4        0.758          62    -1384.        -836.
##  3 Oregon                    4        0.75           92    -1239         -951.
##  4 California                4        0.723         130    -1168.        -834.
##  5 Nevada                    4        0.697          33     -992         -836.
##  6 Nevada                    3        0.531         179     -110          280.
##  7 California                3        0.526         705      -92          277.
##  8 Washington                3        0.511         180      -44.5        322.
##  9 Oregon                    3        0.493         550       33.5        389.
## 10 Arizona                   3        0.488         361       30          276.
## 11 Washington                1        0.333          15      301          431.
## 12 California                1        0.312          80      433          659.
## 13 Oregon                    1        0.298          57      697          695.
## 14 Nevada                    1        0.286          14      396.         667.
## 15 California                2        0.245        1235     1107         1424.
## 16 Arizona                   1        0.245          53      342          714.
## 17 Nevada                    2        0.243         375     1150         1471.
## 18 Washington                2        0.242         330      944         1342.
## 19 Oregon                    2        0.238        1064      963         1414.
## 20 Arizona                   2        0.226         705     1080         1480.
```

```
ggplot(df4,aes(x=customer_state,y=total_claims,fill=Avg_CLV_Loss))+
  geom_tile()+
  scale_fill_viridis_b()+
  labs(title = "Number of Claims vs Customer States",
       subtitle="Loss ratio",
       x="Number of Claims",
       y="States",
       fill="Loss Ratio")
```




```
#-------------------------------------------------------------------------------------------------

# Q5
# Is there a strong correlation between income and monthly premium paid?
# How does it vary across the different marital status and gender?

(df5<- claims_df %>% group_by(marital_status,gender)%>%
    summarise(avg_monthly_premium = mean(monthly_premium),
              n_customers=n(),
              median_monthly_premium=median(monthly_premium),
              max_monthly_premium = max(monthly_premium),
              min_monthly_premium=min(monthly_premium))%>%
    arrange(desc(avg_monthly_premium)))
```

```
## `summarise()` has grouped output by 'marital_status'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 6 x 7
## # Groups:   marital_status [3]
##   marital_status gender avg_monthly_premium n_customers median_monthly_premium
##   <chr>          <chr>                <dbl>       <int>                  <dbl>
## 1 Single         Female                96.2         481                     88
## 2 Married        Female                94.1        2159                     85
## 3 Divorced       Male                  94.1         527                     81
## 4 Divorced       Female                93.8         537                     82
## 5 Single         Male                  93.2         546                     83
## 6 Married        Male                  93.0        1999                     83
## # i 2 more variables: max_monthly_premium <int>, min_monthly_premium <int>
```

```
ggplot(claims_df, aes(x = income, y = monthly_premium, color=gender)) +
  geom_point() +
  facet_wrap(~ marital_status, nrow=3)+
  labs(title = "Income vs. Monthly Premium",
       subtitle="Considering Marital status and gender",
       x="Income",
       y="Monthly Premium",
       color="Gender")+coord_flip()
```

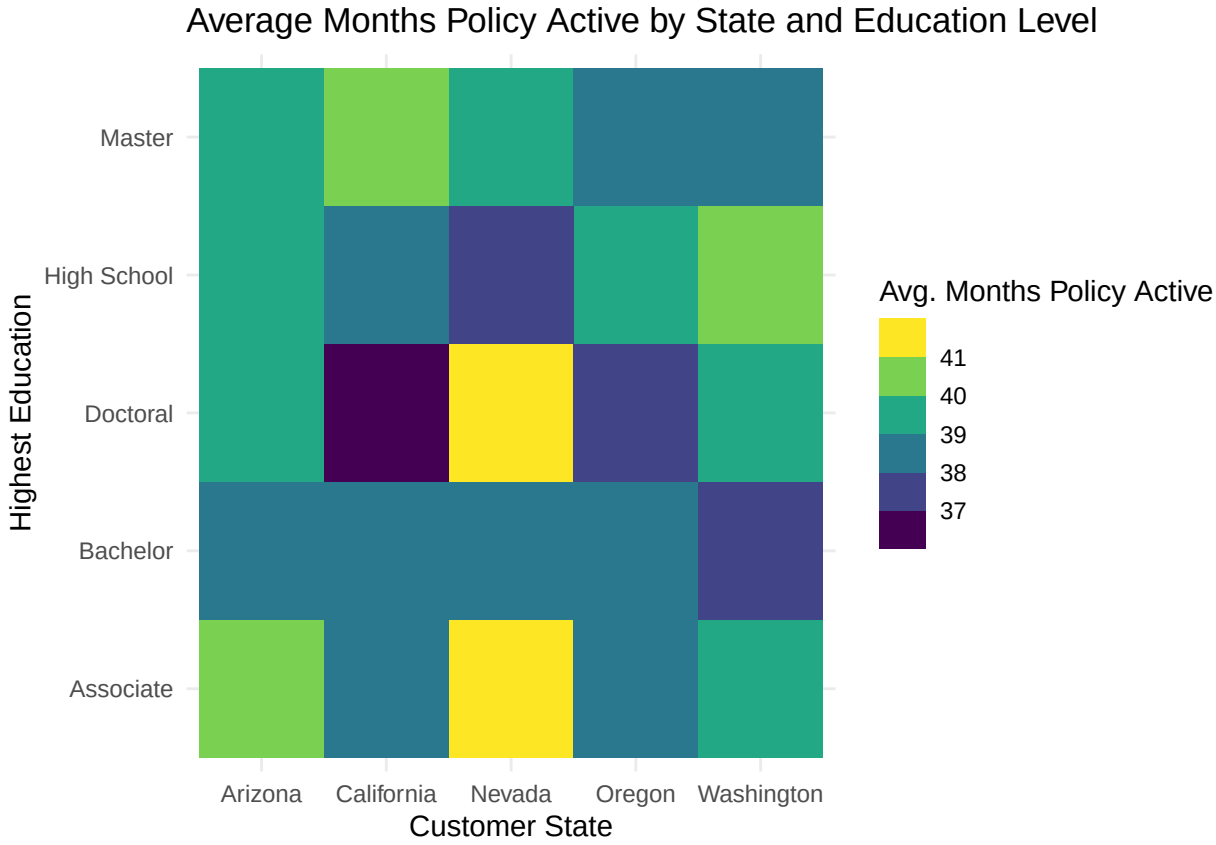

```
#_____________________________________________________________________________________________
# Q6
# What is the relationship between customer state and highest education level in determining
# the average duration of policies active?

(df6 <- claims_df %>%
    group_by(customer_state, highest_education) %>%
    summarise(avg_months_policy_active = mean(months_policy_active)) %>%
    arrange(desc(avg_months_policy_active)))
```

```
## `summarise()` has grouped output by 'customer_state'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 25 x 3
## # Groups:   customer_state [5]
##    customer_state highest_education avg_months_policy_active
##    <chr>          <chr>                                <dbl>
##  1 Nevada         Associate                             42.0
##  2 Nevada         Doctoral                              41.2
```

```
##  3 Washington     High School                    40.6
##  4 California     Master                         40.5
##  5 Arizona        Associate                      40.2
##  6 Arizona        Doctoral                       40.0
##  7 Washington     Associate                      39.7
##  8 Washington     Doctoral                       39.5
##  9 Arizona        Master                         39.5
## 10 Oregon         High School                    39.2
## # i 15 more rows
```

```r
ggplot(df6, aes(x = customer_state, y = highest_education, fill = avg_months_policy_active)) +
  geom_tile() +
  scale_fill_viridis_b() +
  labs(title = "Average Months Policy Active by State and Education Level",
       x = "Customer State",
       y = "Highest Education",
       fill = "Avg. Months Policy Active") +
  theme_minimal()
```

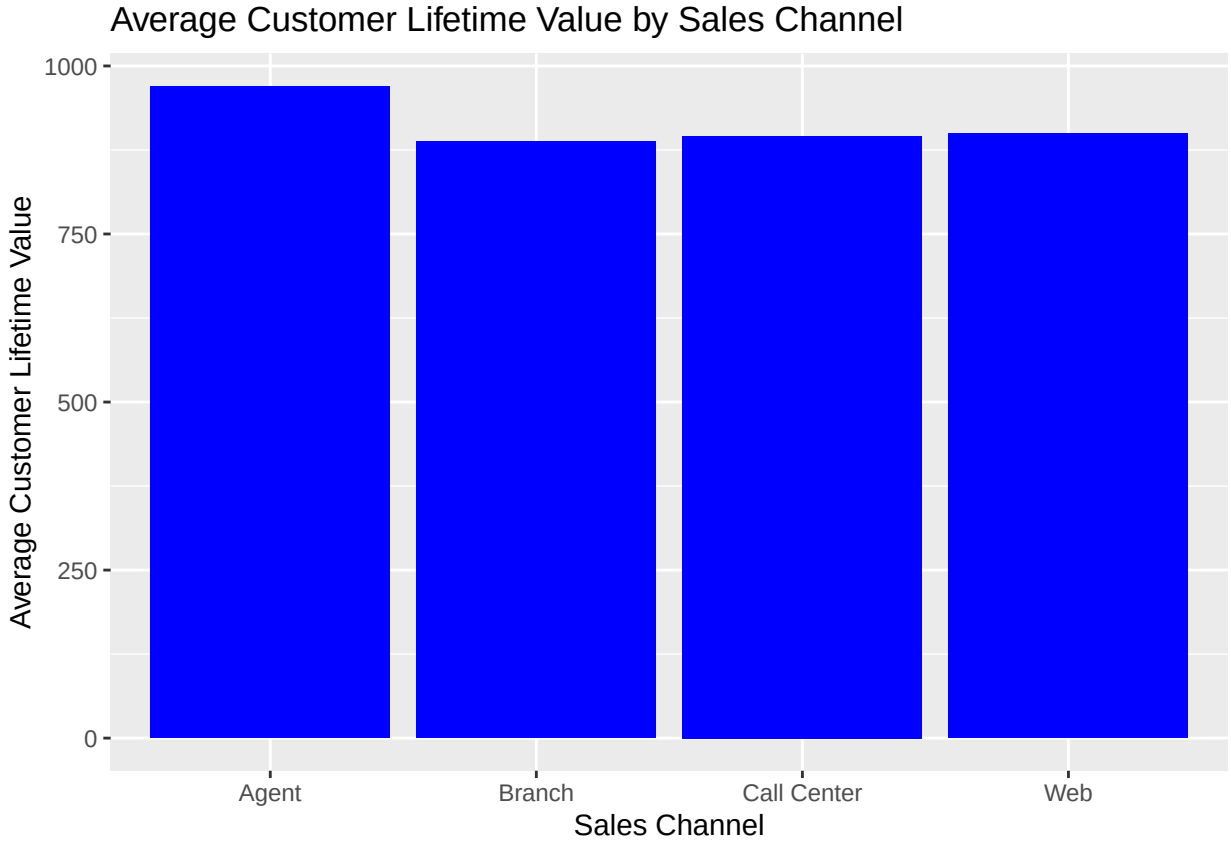



```r
#_____________________________________________________________________________________________
# Q7
# How do different customer acquisition channels impact customer value and claim behavior?
(df7 <- claims_df %>%
    group_by(sales_channel) %>%
    summarise(Average_CLV = mean(customer_lifetime_value),
              Median_CLV = median(customer_lifetime_value),
              Average_Claims = mean(total_claims),
              Average_Claim_Amount = mean(current_claim_amount),
```

```
            Median_Claim_Amount = median(current_claim_amount),
            n = n()) %>% arrange(desc(Average_CLV)))
```

```
## # A tibble: 4 x 7
##   sales_channel Average_CLV Median_CLV Average_Claims Average_Claim_Amount
##   <chr>               <dbl>      <dbl>          <dbl>                <dbl>
## 1 Agent                970.        570           2.38                1623.
## 2 Web                  899.        523           2.41                1602.
## 3 Call Center          896.        612           2.39                1637.
## 4 Branch               888.        594           2.40                1633.
## # i 2 more variables: Median_Claim_Amount <dbl>, n <int>
```

```
ggplot(df7, aes(x = sales_channel, y = Average_CLV)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Average Customer Lifetime Value by Sales Channel",
       x = "Sales Channel",
       y = "Average Customer Lifetime Value")
```




```
#-------------------------------------------------------------------------------------------------

# Q8
# Is there a correlation between customer retention and profitability across all the policy tiers
# offered?
ggplot(claims_df, aes(x = months_policy_active, y = customer_lifetime_value,color=coverage)) +
  geom_point() +
  facet_wrap(~coverage)+
  labs(title = "Customer Retention vs. Profitability",
       x = "Months Policy Active",
```

```
        y = "Customer Lifetime Value",
        color="Policy Tiers")
```