Data Analysis of an Insurance Company Result

```
library(tidyverse)

claims_df <- readRDS(url('https://gmubusinessanalytics.netlify.app/data/claims_df.rds'))
```

Raw Data

```
claims_df
```

```
# A tibble: 6,249 x 20
   custo~1 custo~2 highe~3 emplo~4 gender income resid~5 marit~6 sales~7 cover~8
   <chr>   <fct>   <fct>   <fct>   <fct>   <dbl> <fct>   <fct>   <fct>   <fct>
 1 AA11235 Nevada  Bachel~ Medica~ Female  11167 Suburb~ Married Branch  Basic
 2 AA16582 Washin~ Bachel~ Medica~ Male    14072 Suburb~ Divorc~ Agent   Basic
 3 AA34092 Califo~ Associ~ Employ~ Male    33635 Suburb~ Married Web     Extend~
 4 AA56476 Arizona High S~ Employ~ Female  74454 Suburb~ Single  Call C~ Basic
 5 AA69265 Nevada  Bachel~ Employ~ Female  60817 Suburb~ Single  Web     Premium
 6 AA71604 Arizona Master  Employ~ Female  87560 Suburb~ Married Web     Extend~
 7 AA93585 Califo~ Associ~ Employ~ Male    97024 Urban   Married Branch  Premium
 8 AB21519 Califo~ Associ~ Employ~ Female  93272 Urban   Married Branch  Extend~
 9 AB23825 Califo~ Associ~ Employ~ Male    21509 Suburb~ Single  Agent   Extend~
10 AB26022 Oregon  High S~ Retired Male    26487 Suburb~ Single  Call C~ Basic
# ... with 6,239 more rows, 10 more variables: policy <fct>,
#   vehicle_class <fct>, vehicle_size <fct>, monthly_premium <dbl>,
#   months_policy_active <dbl>, months_since_last_claim <dbl>,
#   current_claim_amount <dbl>, total_claims <dbl>, total_claims_amount <dbl>,
#   customer_lifetime_value <dbl>, and abbreviated variable names
#   1: customer_id, 2: customer_state, 3: highest_education,
#   4: employment_status, 5: residence_type, 6: marital_status, ...
```

Exploratory Data Analysis

Question 1

Which state is responsible for the majority of the profits?

Answer: California has the highest monetary profit at $1898706 which constitues 32.95% of the overall profit. Although the average profit is comparitively lower, it still rakes in the highest profit margin. The total number of claims in California is 5185.

States like Washington and Nevada has the least number of claims.

To add additional R code chunks for your work, select `Insert` then `R` from the top of this notebook file.

```
# Question 01

library(ggplot2)
library(tidyr)
library(dplyr)
```

```
library(tidyverse)
library(skimr)
library(usmap)
claims_df <- readRDS(url('https://gmubusinessanalytics.netlify.app/data/claims_df.rds'))
claims_df = claims_df %>% mutate(revenue = monthly_premium * months_policy_active)
claims_df_1 = claims_df %>% mutate(state = claims_df$customer_state)
profit = sum(claims_df$customer_lifetime_value)
profit
```

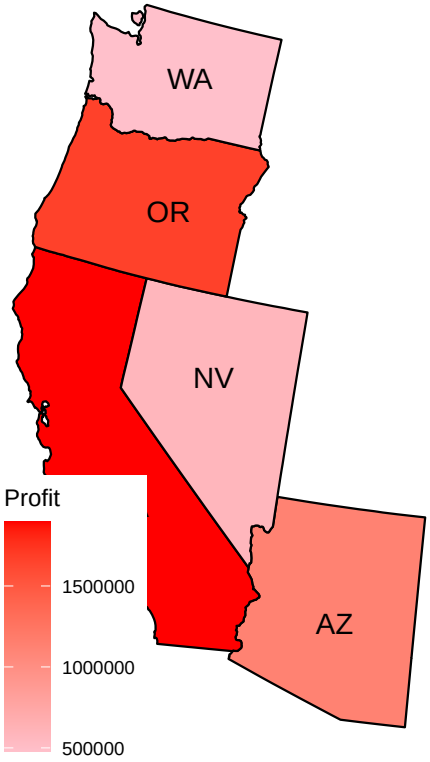
```
[1] 5761975
```

```
q1 = claims_df_1 %>% group_by(state) %>% summarize(
  total_revenue = sum(revenue),
  count = n(),
  total_claims = sum(total_claims),
  sum_amount_claimed = sum(total_claims_amount),
  totalprofit = sum(customer_lifetime_value),
  avg_profit = mean(customer_lifetime_value),
  percentage_profit = (totalprofit/5761975) *100)
q1
```

```
# A tibble: 5 x 8
  state      total_revenue count total_claims sum_amou~1 total~2 avg_p~3 perce~4
  <fct>              <dbl> <int>        <dbl>      <dbl>   <dbl>   <dbl>   <dbl>
1 Washington       2009800   554         1331    1529445  480355    867.    8.34
2 Oregon           6474273  1763         4203    4804195 1670078    947.   29.0
3 California       7864994  2150         5185    5966288 1898706    883.   33.0
4 Nevada           2238504   601         1433    1654854  583650    971.   10.1
5 Arizona          4327041  1181         2794    3197855 1129186    956.   19.6
# ... with abbreviated variable names 1: sum_amount_claimed, 2: totalprofit,
#   3: avg_profit, 4: percentage_profit
```

```
plot_usmap(data = q1,values = "totalprofit",include = c("CA","WA","OR","NV","AZ"), labels = TRUE) +
  scale_fill_gradient(name = "Profit", low = "pink", high =  "red", na.value = "grey50") +labs(title =
```

Statewise customer value



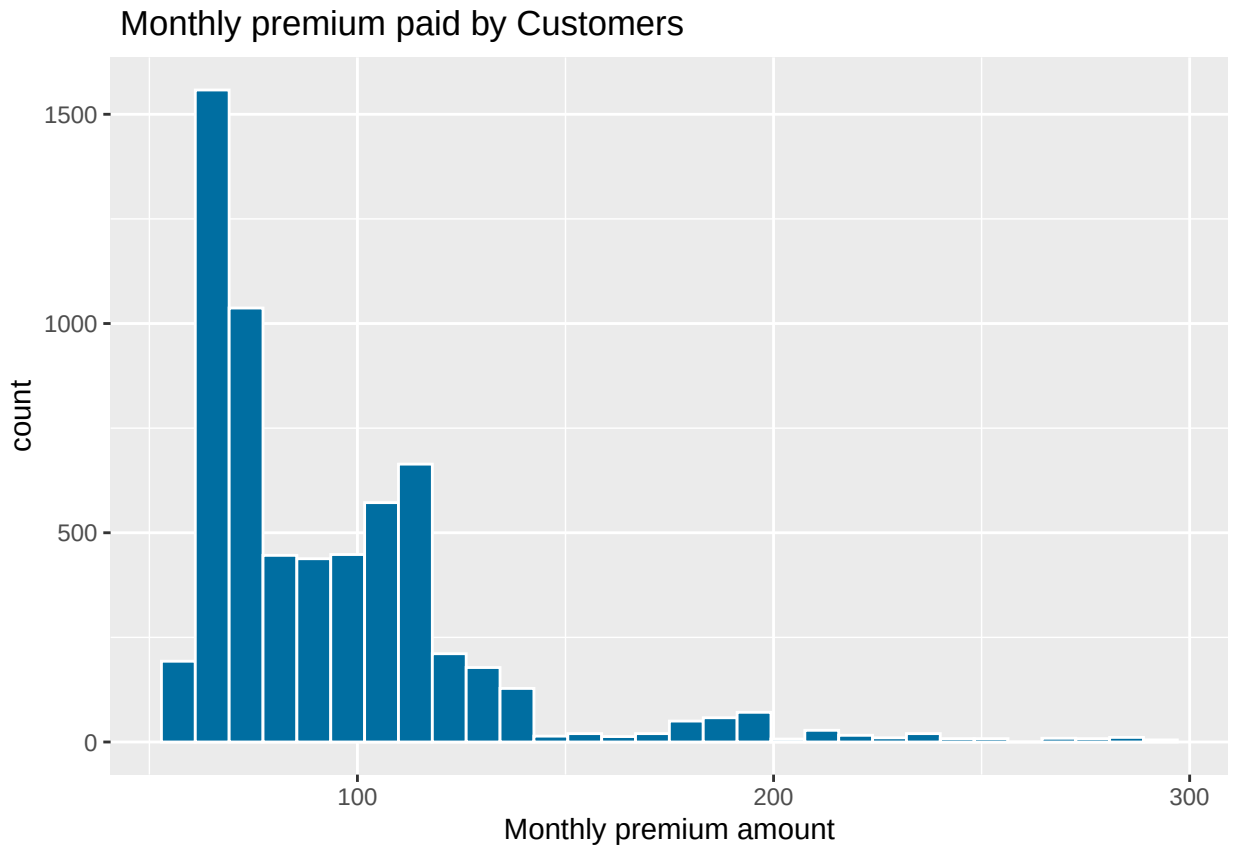
Question 2

How is Monthly premiun paid affect company profits?

Answer:

Customers paying higher monthly premium($100 - $150), although 1835 customers pay in this range. They contribute to an total profit of $3219546 However when you compare this to the most widely paid montly premium in the range of ($50 - $100) by 4052 customers , the profit obtained is $530130.

```
ggplot(claims_df,aes(x = monthly_premium)) + geom_histogram(fill = "#006EA1", color = "white" , bins = 3
  labs(title =" Monthly premium paid by Customers", x = "Monthly premium amount")
```




```
claims_df_updates_3 = claims_df %>% mutate(monthly_premium_range = cut_width(monthly_premium, width = 5
#summary stats


claims_df_updates_3 %>% group_by(monthly_premium_range) %>% summarize(
  count = n(),
  avg_monthly_premium = mean(monthly_premium),
  avg_cust_value = mean(customer_lifetime_value),
  total_profit = sum(customer_lifetime_value)
)
```

```
# A tibble: 5 x 5
  monthly_premium_range count avg_monthly_premium avg_cust_value total_profit
  <fct>                 <int>               <dbl>          <dbl>        <dbl>
1 [50,100]               4052                74.4           131.       530130
2 (100,150]              1835               115.           1755.      3219546
3 (150,200]               232               183.           4572.      1060756
4 (200,250]                92               224.           6198.       570228
5 (250,300]                38               276.          10035.       381315
```

Question 3

Is the profit of the company getting affected by the type of coverage chosen by the customers along with the monthly premium paid?

Answer:

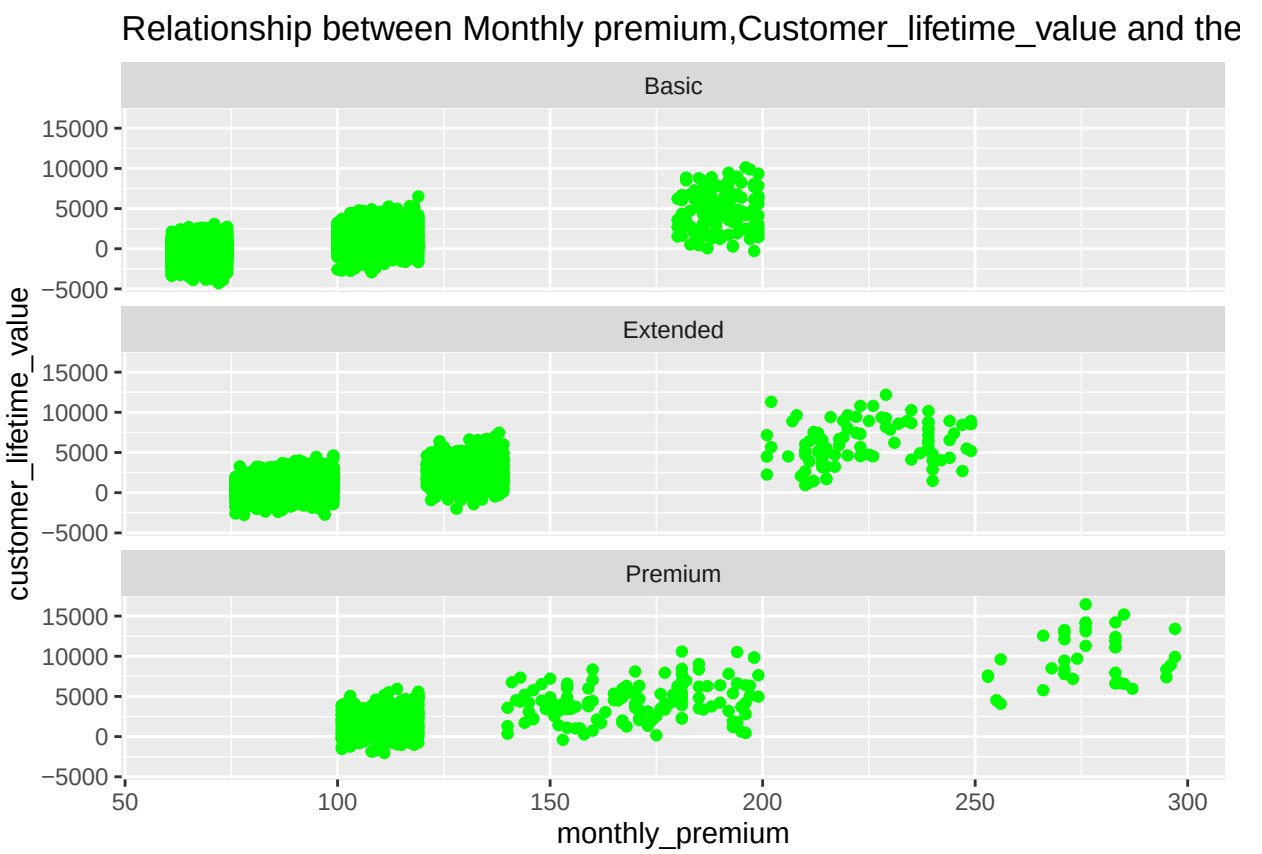
Company faces loss of around 47% in the Basic coverage opted by the 3815 customers paying an average of

$82 per month.Major loss is incurred by the customer choosing basic coverage and paying monthly premium in the range of $61 to $75.

Executive coverage also faces loss of 18% opted by 1858 customer whose average monthly membership is $104.

Premium coverage faces the least loss of 10% opted by 576 customers.

```
ggplot(claims_df, aes(x = monthly_premium, y = customer_lifetime_value)) +
  geom_point(color = "green") +facet_wrap(~coverage ,nrow=3) +labs(title = "Relationship between Monthly
```




```
#Summary needs to be done
claims_df %>% group_by(coverage) %>% summarize(
  avg_monthly_premium = mean(monthly_premium),
  count= n(),
  avg_loss_cust = mean(customer_lifetime_value<0)
)
```

```
# A tibble: 3 x 4
  coverage avg_monthly_premium count avg_loss_cust
  <fct>                  <dbl> <int>         <dbl>
1 Basic                   82.5  3815         0.471
2 Extended               104.   1858         0.189
3 Premium                134.    576         0.108
```

Question 4

How is company's profit affected by the different sales channel?

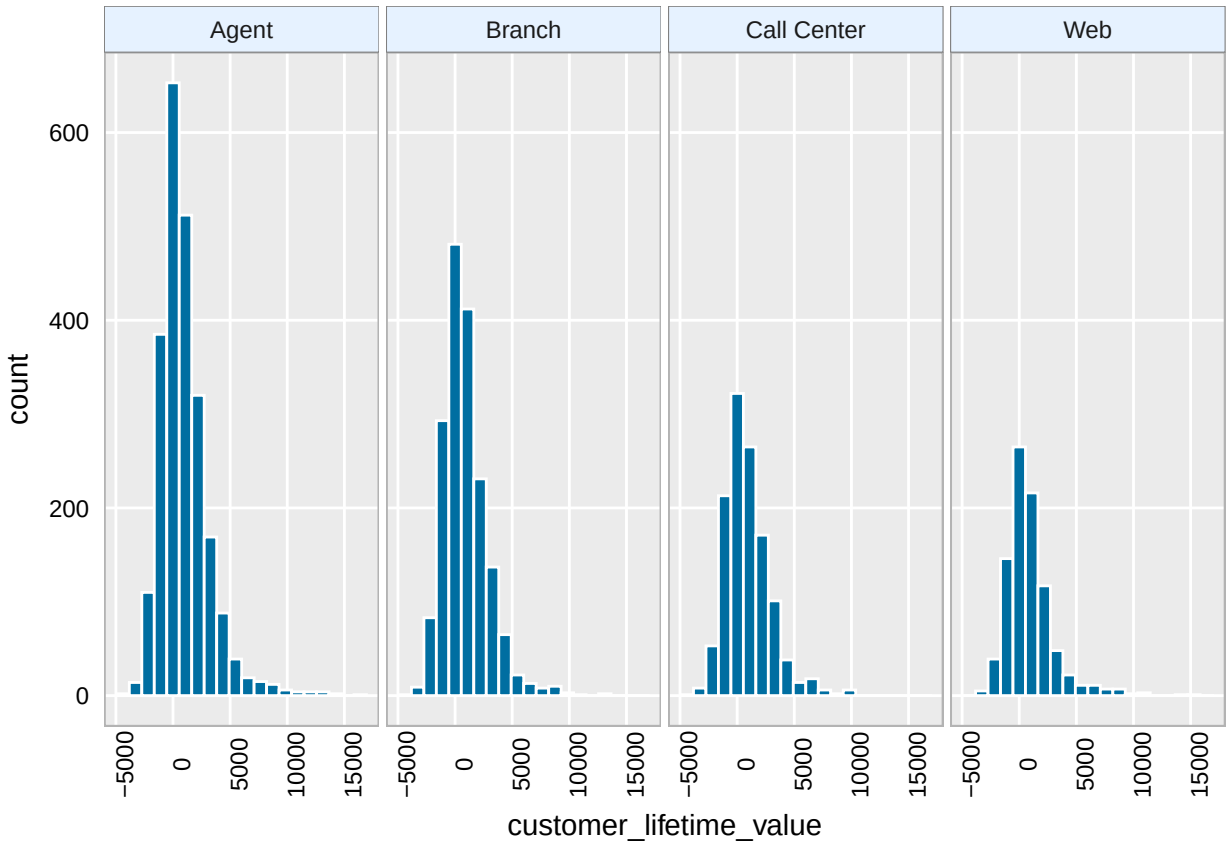
Answer:

Out of 4 sales channel, 2359 customers who took up coverage through third party agents have fetched highest profit of $2288145. Customers joining the company through Branch, call center and web brings around same average profit. Web sales channel has attracted 901 customers which is the least.

```
hw <- theme_gray()+ theme(
  plot.title=element_text(hjust=0.5),
  plot.subtitle=element_text(hjust=0.5),
  plot.caption=element_text(hjust=-.5),
  # strip.text.y = element_blank(),
  strip.background=element_rect(fill=rgb(.9,.95,1),
                                colour=gray(.5), size=.2),
  panel.border=element_rect(fill=FALSE,colour=gray(.70)),
  panel.grid.minor.y = element_blank(),
  panel.grid.minor.x = element_blank(),
  panel.spacing.x = unit(0.10,"cm"),
  panel.spacing.y = unit(0.05,"cm"),
  # axis.ticks.y= element_blank()
  axis.ticks=element_blank(),
  axis.text=element_text(colour="black"),
  axis.text.y=element_text(margin=margin(0,3,0,3)),
  axis.text.x=element_text(margin=margin(-1,0,3,0),angle=90)
)



ggplot(claims_df, aes(x = customer_lifetime_value)) +
  geom_histogram(fill = "#006EA1", color = "white" , bins = 20) +
  facet_grid(~sales_channel) +hw
```




```
#summary stats needs to be made

claims_df %>% group_by(sales_channel) %>% summarize(
  count= n(),
  total_profit = sum(customer_lifetime_value),
  avg_profit = mean(customer_lifetime_value)

)
```

```
# A tibble: 4 x 4
  sales_channel count total_profit avg_profit
  <fct>         <int>        <dbl>      <dbl>
1 Agent          2359      2288145       970.
2 Branch         1771      1572695       888.
3 Call Center    1218      1090921       896.
4 Web             901       810214       899.
```

Question 5

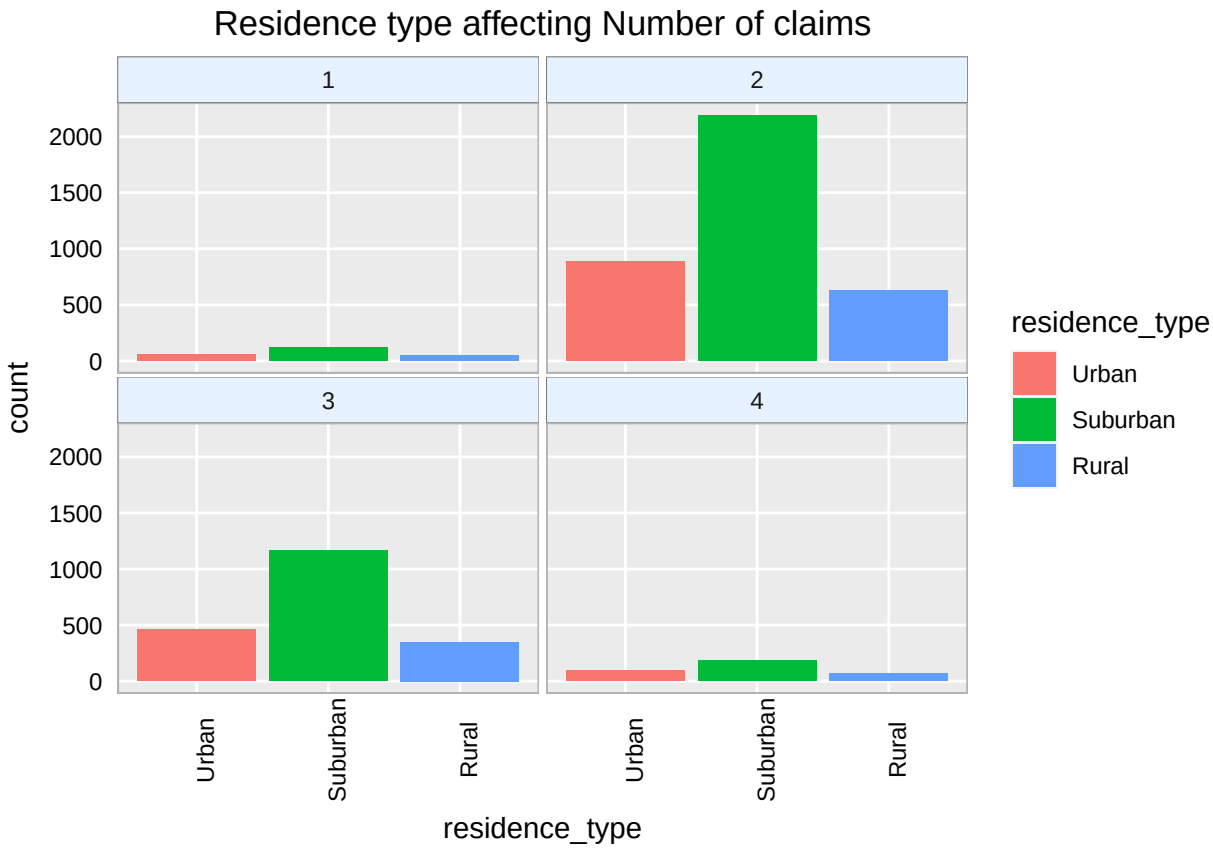
Which residence type has the more more of claims?

Answer:

Sub-Urban has the highest number of 2 times claim and 3 times claim accounting for 2193 and 1165 customers respectively.

Second highest claims are seen in urban areas where the number of 2 times claimed account for 885 customers.

```
ggplot(claims_df, aes( x = residence_type, fill = residence_type)) + geom_bar(stat = "count" ) +
  facet_wrap(~total_claims) +hw +labs(title = "Residence type affecting Number of claims")
```




```
# summary needs to be done

claims_df %>% group_by(residence_type, total_claims) %>% summarize(
  count = n(),
)
```

```
# A tibble: 12 x 3
# Groups:   residence_type [3]
   residence_type total_claims count
   <fct>                 <dbl> <int>
 1 Urban                     1    56
 2 Urban                     2   885
 3 Urban                     3   460
 4 Urban                     4    94
 5 Suburban                  1   118
 6 Suburban                  2  2193
 7 Suburban                  3  1165
 8 Suburban                  4   181
 9 Rural                     1    45
10 Rural                     2   631
11 Rural                     3   350
12 Rural                     4    71
```

Question 6

Which vehicle class fetches higher profit?

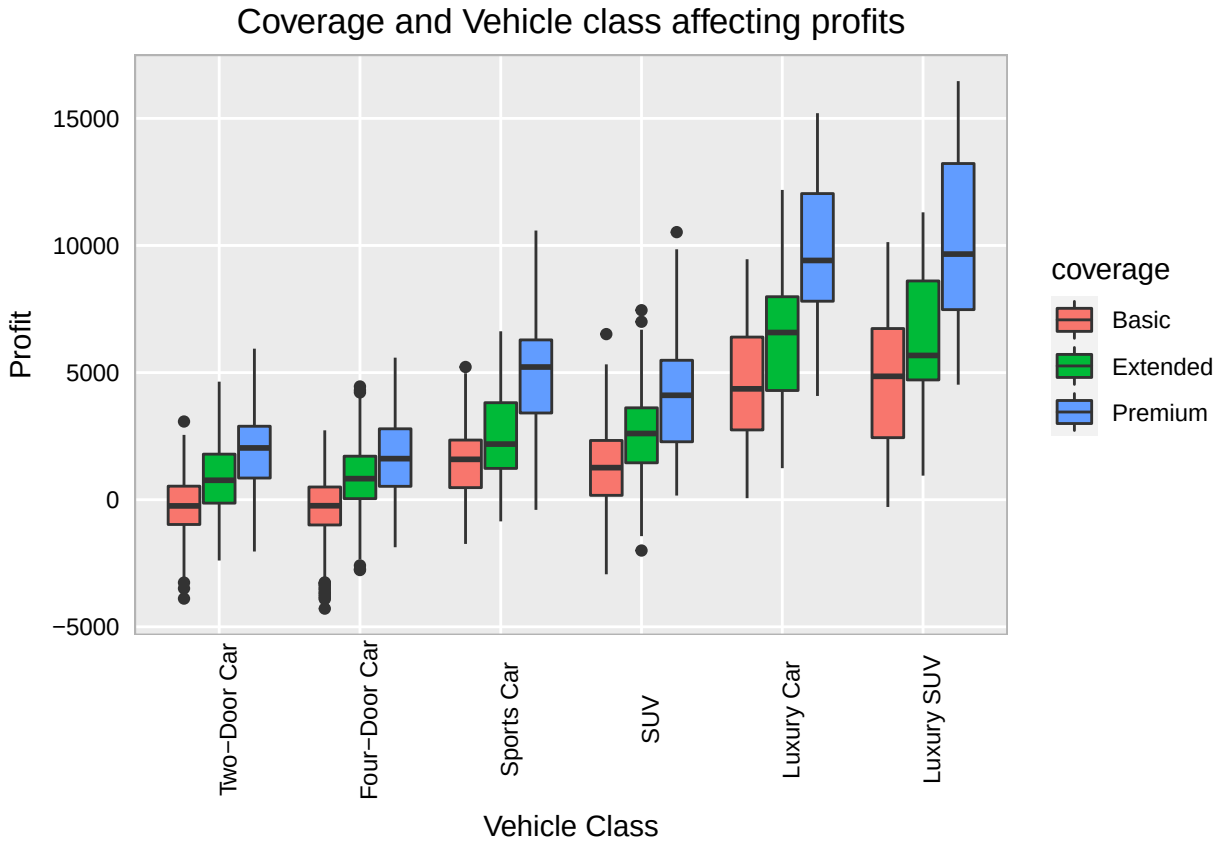
Answer:

Luxury Car has not brought any loss to the company from all the three coverage types. Its average loss is 0%. 2-Door cars and 4-Door cars bring the highest loss having 1292 and 3124 claims respectively.

Customer with 2-Door basic coverage and customers with 4-Door basic coverage bring the highest loss to the company.

Average profit for 2-door car is $269 which is the least among other vehicle class.

```
ggplot(claims_df,aes(x = vehicle_class, y = customer_lifetime_value, fill = coverage)) + geom_boxplot()
  labs(title = "Coverage and Vehicle class affecting profits", x = "Vehicle Class", y = "Profit" ) +hw
```




```
claims_df = claims_df %>% mutate(revenue = monthly_premium * months_policy_active)
# profit of the company increases with the luxury SUV
#2 door cars and 4 door cars bring loss


claims_df %>% group_by(vehicle_class) %>% summarize(
  number_of_claims = n(),
  avg_profit = mean(customer_lifetime_value),
  avg_revenue = mean(revenue),
  loss_occured = any(customer_lifetime_value<0),
  avg_loses = mean(customer_lifetime_value < 0)
)
```

```
# A tibble: 6 x 6
  vehicle_class number_of_claims avg_profit avg_revenue loss_occured avg_loses
  <fct>                   <int>      <dbl>       <dbl> <lgl>            <dbl>
1 Two-Door Car             1292       269.       3026. TRUE           0.459
2 Four-Door Car            3124       271.       3027. TRUE           0.444
3 Sports Car                335      2159.       4861. TRUE           0.110
4 SUV                      1246      1861.       4601. TRUE           0.154
5 Luxury Car                119      5670.       8362. FALSE          0
6 Luxury SUV                133      6382.       8966. TRUE           0.00752
```

Question 7

Which policies offers higher profit to the company?

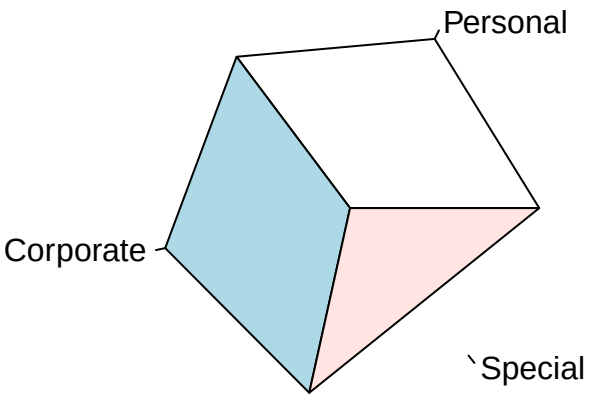
Answer:

4658 Customers have opted for personal policy which fetched a profit of $4302430. Special policies are least bought by the customers whose average profit($745.9582) is less compared to other policies.

```
q4 = claims_df %>% group_by(policy) %>% summarize(
  avg_profit = mean(customer_lifetime_value),
  count = n(),
  total_profit = sum(customer_lifetime_value),
  med_profot = median(customer_lifetime_value)
)
q4
```

```
# A tibble: 3 x 5
  policy    avg_profit count total_profit med_profot
  <fct>          <dbl> <int>        <dbl>      <dbl>
1 Personal        924.  4658      4302430       560
2 Corporate       951.  1328      1263358       632.
3 Special         746.   263       196187       432
```

```
pie(q4$avg_profit , labels = c("Personal","Corporate","Special") , edges=10 , main ="Profit through dif
    )
```



Question 8

What is the relationship between the coverage plan chosen by the customer and their income?

Answer:

Customers with income in the range of $20000 to $40000 choose Basic coverage over other plans.

```
claims_df_updates_q8 = claims_df %>% mutate(income_range = cut_width(income, width = 20000, boundary = 0

ggplot(claims_df_updates_q8, aes( x = income_range )) +
  geom_bar(stat = "count", color = "blue") +hw +labs(title ="Income of the customer and the coverage op
```