Stat-295-HW2

Sabahattin Alp Kocabaş, Başak Kabaloğlu, Efe Örencik

2024-05-26

Question 1

- **Loading necessary packages.**

```
library(ggplot2)
library(tidyverse)
```

- **Reading the data.**

```
data <- read.csv("social_network_ad.csv")
```

- **Displaying the structure and summary of the data set.**

```
str(data)
```

```
## 'data.frame':    400 obs. of  7 variables:
##  $ X              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ User.ID        : int  15624510 15810944 15668575 15603246 15804002 15728773 15598
044 15694829 15600575 15727311 ...
##  $ Gender         : chr  "Male" "Male" "Female" "Female" ...
##  $ Age            : int  19 35 26 27 19 27 27 32 25 35 ...
##  $ EstimatedSalary: int  19000 20000 43000 57000 76000 58000 84000 150000 33000 6500
0 ...
##  $ Purchased      : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ GiftTicket     : int  0 1 0 0 0 0 0 1 0 1 ...
```

```
summary(data)
```

```
##        X            User.ID           Gender               Age
##  Min.   :  1.0   Min.   :15566689   Length:400         Min.   :18.00
##  1st Qu.:100.8   1st Qu.:15626764   Class :character   1st Qu.:29.75
##  Median :200.5   Median :15694342   Mode  :character   Median :37.00
##  Mean   :200.5   Mean   :15691540                      Mean   :37.66
##  3rd Qu.:300.2   3rd Qu.:15750363                      3rd Qu.:46.00
##  Max.   :400.0   Max.   :15815236                      Max.   :60.00
##  EstimatedSalary    Purchased        GiftTicket
##  Min.   : 15000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 43000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 70000   Median :0.0000   Median :1.0000
##  Mean   : 69743   Mean   :0.3575   Mean   :0.7225
##  3rd Qu.: 88000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :150000   Max.   :1.0000   Max.   :1.0000
```

```
head(data)
```

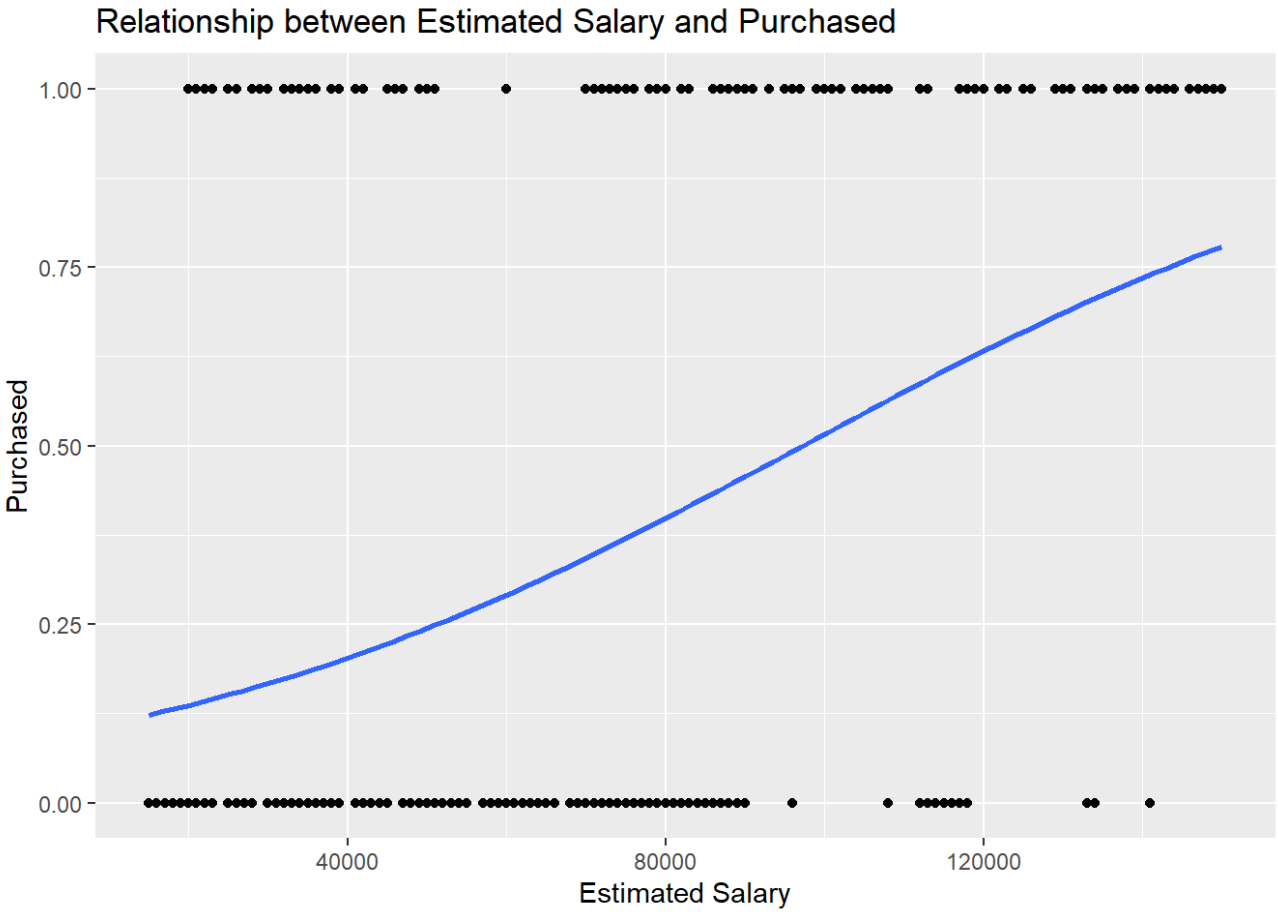
```
##   X  User.ID Gender Age EstimatedSalary Purchased GiftTicket
## 1 1 15624510   Male  19           19000         0          0
## 2 2 15810944   Male  35           20000         0          1
## 3 3 15668575 Female  26           43000         0          0
## 4 4 15603246 Female  27           57000         0          0
## 5 5 15804002   Male  19           76000         0          0
## 6 6 15728773   Male  27           58000         0          0
```

- **Our observations for the dataset**

  The dataset has no missing values. The average age is approximately 37.66 years with a standard deviation of 10.48 years. Estimated salaries range from $15,000 to $150,000, with an average of $69,742.50. About 35.75% of individuals made a purchase. Approximately 72.25% received a gift ticket.

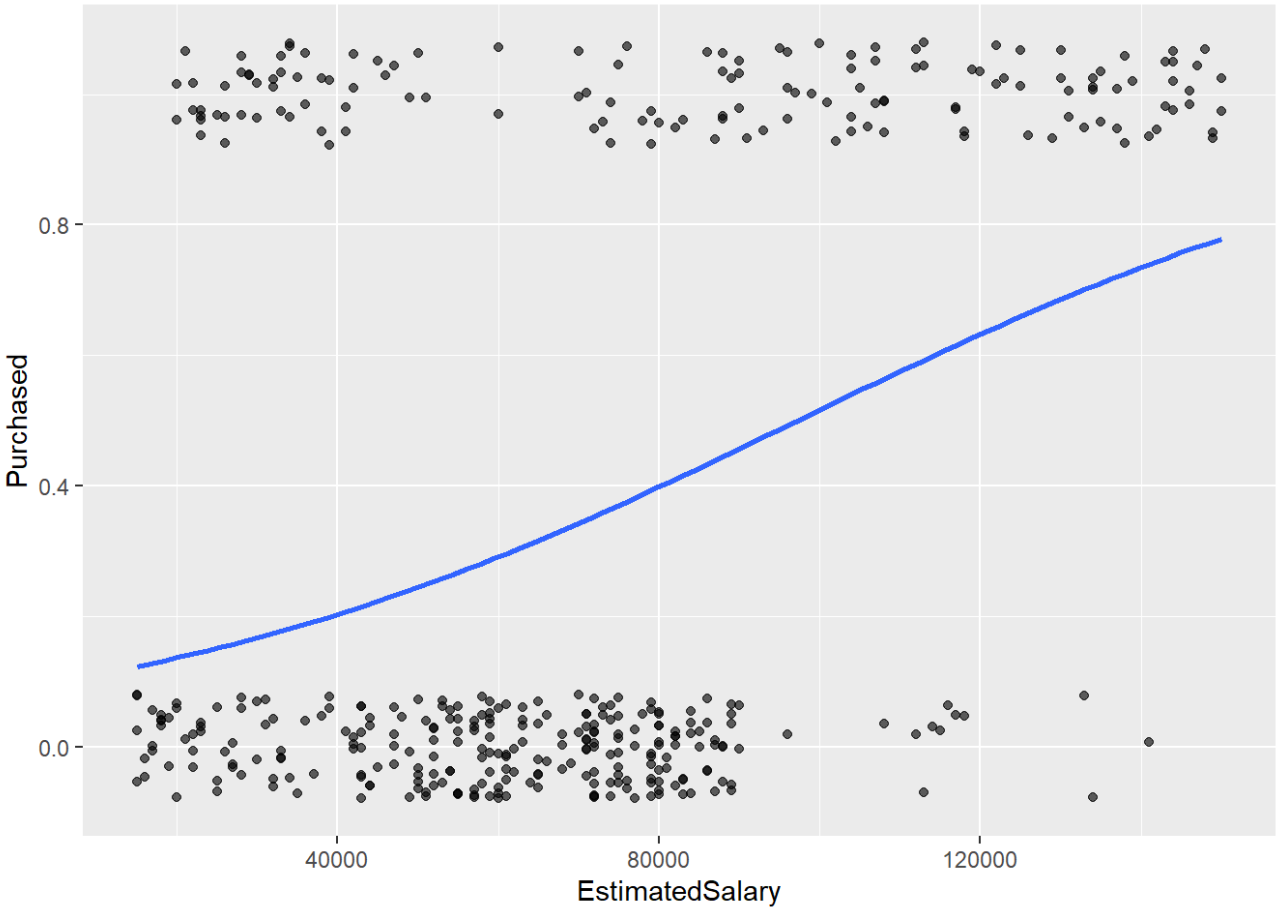
- **(i) Plot showing the relationship between Estimated Salary and Purchased.**

```
ggplot(data, aes(x =EstimatedSalary, y= Purchased)) +
  geom_point() +
  labs(title = "Relationship between Estimated Salary and Purchased",
       x = "Estimated Salary",
       y = "Purchased")+
  geom_smooth(method = "glm", se=FALSE, method.args = list(family = "binomial"))
```



- **(i) Using geom_jitter() to make our graph more informative.**

```
ggplot(data, aes(x = EstimatedSalary, y= Purchased)) +
  geom_jitter(width = 0.5, height =0.08,alpha=0.6) +
  geom_smooth(method = "glm", se=FALSE, method.args = list(family="binomial"))
```



- **(ii) Fitting the logistic regression model.**

```
logit<- glm(Purchased ~ EstimatedSalary, data= data, family = binomial)
summary(logit)
```

```
##
## Call:
## glm(formula = Purchased ~ EstimatedSalary, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6403  -0.9250  -0.6955   0.9851   1.9959
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.323e+00  2.855e-01  -8.134 4.14e-16 ***
## EstimatedSalary  2.388e-05  3.516e-06   6.790 1.12e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 521.57  on 399  degrees of freedom
## Residual deviance: 467.73  on 398  degrees of freedom
## AIC: 471.73
##
## Number of Fisher Scoring iterations: 4
```

- **(ii) Model Interpretation.**

The logistic regression model estimates the relationship between EstimatedSalary and the probability of purchasing a product. The summary provides coefficients that can be interpreted as follows:

Intercept (β0): The log-odds of purchasing when EstimatedSalary is zero

EstimatedSalary (β1): The change in log-odds of purchasing for a one-unit increase in EstimatedSalary.

- **(ii) Odds Ratio and Our Comment.**

```
odds_ratio <- exp(coef(logit))
odds_ratio
```

```
##     (Intercept) EstimatedSalary
##      0.09800858      1.00002388
```

The odds of purchasing increase with increasing estimated salary.

- **(iii) Calculating predicted probability with an estimated salary of $22,000.**

```
salary <- data.frame(EstimatedSalary = 22000)
predicted_prob <- predict(logit, salary, type = "response")
predicted_prob
```

```
##         1
## 0.1421615
```

The predicted probability of purchasing a product for an individual with an estimated salary of $22,000 is approximately 0.1421615.

- **(iv) Converting the Gender column into binary variables (0 for female and 1 for male), and model of Gender.**

```
data$GenderBinary <- ifelse(data$Gender == "Male", 1, 0)

logit_gender <- glm(Purchased ~ GenderBinary, family = binomial,data = data)
summary(logit_gender)
```

```
##
## Call:
## glm(formula = Purchased ~ GenderBinary, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9736  -0.9736  -0.9062   1.3959   1.4754
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.5004     0.1444  -3.464 0.000531 ***
## GenderBinary  -0.1775     0.2091  -0.849 0.395858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 521.57  on 399  degrees of freedom
## Residual deviance: 520.85  on 398  degrees of freedom
## AIC: 524.85
##
## Number of Fisher Scoring iterations: 4
```

- **(iv) Odds Ratio.**

```
odds_ratios <- exp(coef(logit_gender))
odds_ratios
```

```
##  (Intercept) GenderBinary
##    0.6062992    0.8373626
```

The odds ratio tells us how the odds of purchasing change for males compared to females. Since the odds ratio is less than 1, it suggests that males have a lower probability of purchasing compared to females.

- **(v) New model to assess the impact of having a gift ticket on the likelihood of purchasing a product.**

```
logit_gift <- glm(Purchased ~  GiftTicket, data = data, family = binomial)

summary(logit_gift)
```

```
##
## Call:
## glm(formula = Purchased ~ GiftTicket, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1336  -1.1336  -0.3334   1.2218   2.4157
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8622     0.4197  -6.819 9.17e-12 ***
## GiftTicket    2.7583     0.4360   6.327 2.50e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 521.57  on 399  degrees of freedom
## Residual deviance: 446.54  on 398  degrees of freedom
## AIC: 450.54
##
## Number of Fisher Scoring iterations: 5
```

- **(v) Odds Ratio for GiftTicket.**

```
odds_ratio_gift <- exp(coef(logit_gift))
odds_ratio_gift
```

```
## (Intercept)  GiftTicket
##  0.05714286 15.77302629
```

Since the odds ratio is greater than 1, it suggests that having a gift ticket increases the odds of purchasing.

- **(vi) Computing the predicted probability of purchasing a product for each level of the GiftTicket variable.**

```
predictiondata <- data.frame(GiftTicket = c(0, 1))

predictiondata$predicted_prob <- predict(logit_gift, newdata = predictiondata, type =
"response")

print(predictiondata)
```

```
##   GiftTicket predicted_prob
## 1          0     0.05405405
## 2          1     0.47404844
```

Question 2

- **Our Plan.**

For this question we chose the given airbnb dataset. We wanted our application to have three tabs,

1-Summary Statistics,

2-Interactive NYC Map,

3-Filtered List.

We made a brief comment for Summary statistics, and created a filtering system for Neighborhood Group, Room Type, Price Range variables for the interactive NYC Map. In addition to the filter containing the same variables for the Filtered List tab, we added another filter where we can select the columns we want to see in the list suitable for filtering.

- **Loading necessary packages.**

```
library(shiny)
library(leaflet)
library(dplyr)
```

- **Load the data set.**

```
data <- read.csv("AB_NYC_2019.csv")
```

- **Defining UI.**

```
ui <- navbarPage("Airbnb Listings for NYC",

  # Summary Statistics Tab
  tabPanel("Summary Statistics",
           fluidPage(titlePanel("Summary Statistics"),
             mainPanel(verbatimTextOutput("summary"),
               h3("Comments:"),
               print("This table contains summary statistics for
                     various features of Airbnb listings. The dataset
                     includes a total of 48,895 records. Prices vary widely,
                     ranging from a minimum of 0 to a maximum of 10,000 USD.
                     The average price is 152.7 USD, but the median price is
                     106 USD,indicating a right-skewed distribution."))),

  # Interactive NYC Map Tab
  tabPanel("Interactive NYC Map",
           fluidPage(titlePanel("NYC Map"),
             sidebarLayout(sidebarPanel(selectInput("neighborhood_group", "Neighborhood
Group:",
                                choices = unique(data$neighbourhood_group),
                                selected = "Manhattan"),
                 selectInput("room_type", "Room Type:",
                                choices = unique(data$room_type),
                                selected = "Entire home/apt"),
                 sliderInput("price", "Price Range:",
                                min = min(data$price), max = max(data$price),
                                value = c(min(data$price), max(data$price)))),
               mainPanel(leafletOutput("map"))))),

  # Filtered List Tab
  tabPanel("Filtered List",
           fluidPage(titlePanel("Filtered List"),
             sidebarLayout(sidebarPanel(selectInput("neighborhood_group_list", "Neighbo
rhood Group:",
                                choices = unique(data$neighbourhood_group),
                                selected = "Manhattan"),
                 selectInput("room_type_list", "Room Type:",
                                choices = unique(data$room_type),
                                selected = "Entire home/apt"),
                 sliderInput("price_list", "Price Range:",
                                min = min(data$price), max = max(data$price),
                                value = c(min(data$price), max(data$price))),
                 checkboxGroupInput("columns", "Select Columns to Display:",
                                    choices = names(data),
                                    selected = names(data))),
               mainPanel(tableOutput("filtered_list"))))))
```

```
## [1] "This table contains summary statistics for\n                    various featur
es of Airbnb listings. The dataset \n                      includes a total of 48,895 re
cords. Prices vary widely,\n                     ranging from a minimum of 0 to a maxim
um of 10,000 USD.\n                     The average price is 152.7 USD, but the median
price is \n                      106 USD,indicating a right-skewed distribution."
```

- **Defining server logic.**

```
server <- function(input, output, session) {

  # Summary Statistics of data
  output$summary <- renderPrint({summary(data)})

  # Interactive NYC Map
  filtered_data <- reactive({data %>%
      filter(neighbourhood_group == input$neighborhood_group,
             room_type == input$room_type,
             price >= input$price[1],
             price <= input$price[2])})

  output$map <- renderLeaflet({leaflet(filtered_data()) %>%
      addTiles() %>%
      addCircleMarkers(~longitude, ~latitude,
                       popup = ~paste(name, "<br>", "Price: $", price, "<br>", "Room Ty
pe: ", room_type),
                       radius = 3, color = "blue", stroke = FALSE, fillOpacity = 0.7)})

  # Filtered List
  filtered_list_data <- reactive({data %>%
      filter(neighbourhood_group == input$neighborhood_group_list,
             room_type == input$room_type_list,
             price >= input$price_list[1],
             price <= input$price_list[2]) %>%
      select(all_of(input$columns))})

  output$filtered_list <- renderTable({
    filtered_list_data()}, rownames = TRUE)}
```

- **Running the app.**

```
shinyApp(ui = ui, server = server)
```

Shiny applications not supported in static R Markdown documents