Assignment 2

Assignment 2 code

```
--------------------------------------------------------------------------------

# Load necessary libraries

library(sqldf)

# Read data from CSV file

file_path <- "C:/Users/Alex/Desktop/OnlineBookClub.csv"

customer_data <- read.csv(file_path)

# Check the structure of the data

print(str(customer_data))

customer_data$subscribe <- ifelse(customer_data$subscribe == "yes", 1, ifelse(customer_data$subscribe == "no", 0, NA))

# Check for missing values in 'subscribe' variable

missing_values <- sum(is.na(customer_data$subscribe))

if (missing_values > 0) {

  # Handle missing values by removing corresponding rows or imputing them

  # For simplicity, we'll remove rows with missing 'subscribe' values

  customer_data <- customer_data[!is.na(customer_data$subscribe), ]

}

# Verify 'subscribe' variable after transformation

print(table(customer_data$subscribe))


# SQL Analysis

# 1. Basic statistics

basic_stats <- sqldf("

  SELECT

    AVG(total_) AS avg_total_spent,

    STDEV(total_) AS std_total_spent,

    AVG(purch) AS avg_total_books_purchased,

    STDEV(purch) AS std_total_books_purchased,

    AVG(last) AS avg_months_since_last_purchase,
```

```
    STDEV(last) AS std_months_since_last_purchase
  FROM
    customer_data
")
# Print basic statistics
print(basic_stats)
# 2. Total number subscribed to the online book club by gender
club_subscription <- sqldf("
  SELECT
    gender,
    COUNT(*) AS total_subscribers
  FROM
    customer_data
  WHERE
    subscribe = 'yes'
  GROUP BY
    gender
")
# Print club subscription by gender
print(club_subscription)

# 3. Percentage of females and males subscribed to the book club
club_subscription_percent <- sqldf("
  SELECT
    gender,
    COUNT(*) * 100.0 / (SELECT COUNT(*) FROM customer_data WHERE gender = 'F') AS percentage_subscribed
  FROM
    customer_data
  WHERE
    subscribe = 'yes'
  GROUP BY
    gender
")
```

```
# Print percentage of females and males subscribed to the book club

print(club_subscription_percent)

# Regression Models

# Create IsFemale variable based on gender

customer_data$IsFemale <- ifelse(customer_data$gender == "F", 1, 0)

# Linear regression model

linear_model <- lm(total_ ~ IsFemale + first + purch + child + youth + cook + do_it + refernce + art + geog, data =
customer_data)

# Print summary of linear regression model

print(summary(linear_model))

# Logistic regression model

logistic_model <- glm(subscribe ~ last + total_ + IsFemale + purch + child + youth + cook + do_it + refernce + art +
geog, data = customer_data, family = binomial)

# Print summary of logistic regression model

print(summary(logistic_model))
```

---

1. **Months Since Last Purchase (last)**:

   - Interpretation: For each additional month since the customer's last purchase in offline stores, the log odds of subscribing to the online book club decrease by approximately 0.0947.

   - Recommendation: Target customers who have made recent purchases in offline stores with promotional offers to encourage them to subscribe to the online book club.

2. **Total Dollars Spent (total_)**:

   - Interpretation: For each additional dollar spent by the customer in offline stores, the log odds of subscribing to the online book club increase by approximately 0.0011.

   - Recommendation: Encourage higher spending in offline stores by offering incentives or rewards tied to online subscriptions.

3. Gender (IsFemale)**:

   - Interpretation: Female customers have lower log odds of subscribing to the online book club compared to male customers, with a coefficient of -0.7607.

   - Recommendation: Implement targeted marketing campaigns aimed at female customers to increase their engagement and subscription rates.

4. **Number of Books Purchased (purch)**:

   - Interpretation: For each additional book purchased by the customer in offline stores, the log odds of subscribing to the online book club increase by approximately 0.5743.

- Recommendation: Promote the online book club as a value-added service for customers who frequently purchase books offline, incentivizing them to subscribe for access to exclusive content or discounts.

5. **Genre Preferences (child, youth, cook, do_it, refernce, art)**:

- Interpretation: Customers who purchase specific genres of books have varying effects on their likelihood of subscribing to the online book club.

- Recommendation: Tailor promotional content and offerings based on customers' genre preferences to appeal to their interests and increase subscription rates.

OUTPUT OF CODE

```
# Load necessary libraries
> library(sqldf)
>
> # Read data from CSV file
> file_path <- "C:/Users/Alex/Desktop/OnlineBookClub.csv"  # Specify the full
file path
> customer_data <- read.csv(file_path)
>
> # Check the structure of the data
> print(str(customer_data))
'data.frame':  50000 obs. of  16 variables:
 $ acctnum  : int  10001 10002 10003 10004 10005 10006 10007 10008 10009 10010
...
 $ gender   : chr  "M/NB" "M/NB" "F" "F" ...
 $ first    : int  49 39 19 7 15 7 25 41 65 11 ...
 $ last     : int  29 27 15 7 15 7 25 1 5 11 ...
 $ book_    : int  109 35 25 15 15 15 15 124 130 15 ...
 $ nonbook_ : int  248 103 147 257 134 98 0 114 288 108 ...
 $ total_   : int  357 138 172 272 149 113 15 238 418 123 ...
 $ purch    : int  10 3 2 1 1 1 1 11 11 1 ...
 $ child    : int  3 0 0 0 0 0 0 2 0 0 ...
 $ youth    : int  2 1 0 0 0 1 0 1 2 1 ...
 $ cook     : int  2 0 2 0 1 0 0 2 3 0 ...
 $ do_it    : int  0 1 0 0 0 0 1 3 2 0 ...
 $ refernce : int  1 0 0 1 0 0 0 0 0 0 ...
 $ art      : int  0 0 0 0 0 0 0 0 3 0 ...
 $ geog     : int  2 1 0 0 0 0 0 3 1 0 ...
 $ subscribe: chr  "no" "no" "no" "no" ...
NULL
>
> # Ensure 'subscribe' variable contains only 'yes' and 'no' values
> # If it contains other values, handle them appropriately
```

```
> customer_data$subscribe <- ifelse(customer_data$subscribe == "yes", 1,
ifelse(customer_data$subscribe == "no", 0, NA))
>
> # Check for missing values in 'subscribe' variable
> missing_values <- sum(is.na(customer_data$subscribe))
>
> if (missing_values > 0) {
+   # Handle missing values by removing corresponding rows or imputing them
+   # For simplicity, we'll remove rows with missing 'subscribe' values
+   customer_data <- customer_data[!is.na(customer_data$subscribe), ]
+ }
>
> # Verify 'subscribe' variable after transformation
> print(table(customer_data$subscribe))

    0     1
45478  4522
>
> # SQL Analysis
> # 1. Basic statistics
> basic_stats <- sqldf("
+   SELECT
+     AVG(total_) AS avg_total_spent,
+     STDEV(total_) AS std_total_spent,
+     AVG(purch) AS avg_total_books_purchased,
+     STDEV(purch) AS std_total_books_purchased,
+     AVG(last) AS avg_months_since_last_purchase,
+     STDEV(last) AS std_months_since_last_purchase
+   FROM
+     customer_data
+ ")
>
> # Print basic statistics
> print(basic_stats)
  avg_total_spent std_total_spent avg_total_books_purchased
std_total_books_purchased
1        208.3183        101.3573                   3.89022
3.47627
  avg_months_since_last_purchase std_months_since_last_purchase
1                       12.35816                       8.153091
>
> # 2. Total number subscribed to the online book club by gender
> club_subscription <- sqldf("
+   SELECT
+     gender,
+     COUNT(*) AS total_subscribers
+   FROM
+     customer_data
```

```
+   WHERE
+     subscribe = 'yes'
+   GROUP BY
+     gender
+ ")
>
> # Print club subscription by gender
> print(club_subscription)
[1] gender            total_subscribers
<0 rows> (or 0-length row.names)
>
> # 3. Percentage of females and males subscribed to the book club
> club_subscription_percent <- sqldf("
+   SELECT
+     gender,
+     COUNT(*) * 100.0 / (SELECT COUNT(*) FROM customer_data WHERE gender =
'F') AS percentage_subscribed
+   FROM
+     customer_data
+   WHERE
+     subscribe = 'yes'
+   GROUP BY
+     gender
+ ")
>
> # Print percentage of females and males subscribed to the book club
> print(club_subscription_percent)
[1] gender                percentage_subscribed
<0 rows> (or 0-length row.names)
>
> # Regression Models
> # Create IsFemale variable based on gender
> customer_data$IsFemale <- ifelse(customer_data$gender == "F", 1, 0)
>
> # Linear regression model
> linear_model <- lm(total_ ~ IsFemale + first + purch + child + youth + cook
+ do_it + refernce + art + geog, data = customer_data)
>
> # Print summary of linear regression model
> print(summary(linear_model))

Call:
lm(formula = total_ ~ IsFemale + first + purch + child + youth +
    cook + do_it + refernce + art + geog, data = customer_data)

Residuals:
     Min       1Q   Median       3Q      Max
-153.021  -75.383   -0.288   75.390  151.728
```

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.62753    0.95088 157.358   <2e-16 ***
IsFemale      0.67856    0.86483   0.785    0.433
first        -0.03351    0.03831  -0.874    0.382
purch        15.18183    0.52843  28.730   <2e-16 ***
child         0.08238    0.67390   0.122    0.903
youth         0.23590    0.81773   0.288    0.773
cook          0.47399    0.66507   0.713    0.476
do_it        -0.17382    0.78547  -0.221    0.825
refernce     -0.67025    0.86793  -0.772    0.440
art          -0.72179    0.81794  -0.882    0.378
geog               NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86.87 on 49990 degrees of freedom
Multiple R-squared:  0.2656,   Adjusted R-squared:  0.2655
F-statistic:  2009 on 9 and 49990 DF,  p-value: < 2.2e-16

>
> # Logistic regression model
> logistic_model <- glm(subscribe ~ last + total_ + IsFemale + purch + child +
youth + cook + do_it + refernce + art + geog, data = customer_data, family =
binomial)
>
> # Print summary of logistic regression model
> print(summary(logistic_model))

Call:
glm(formula = subscribe ~ last + total_ + IsFemale + purch +
    child + youth + cook + do_it + refernce + art + geog, family = binomial,
    data = customer_data)

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.6001096  0.0520980  -30.71  < 2e-16 ***
last        -0.0947124  0.0027924  -33.92  < 2e-16 ***
total_       0.0011160  0.0001982    5.63  1.8e-08 ***
IsFemale    -0.7607204  0.0357608  -21.27  < 2e-16 ***
purch        0.5742763  0.0186311   30.82  < 2e-16 ***
child       -0.7604925  0.0271900  -27.97  < 2e-16 ***
youth       -0.6872508  0.0333766  -20.59  < 2e-16 ***
cook        -0.8445973  0.0273460  -30.89  < 2e-16 ***
do_it       -1.1134410  0.0346660  -32.12  < 2e-16 ***
refernce    -0.3395887  0.0329918  -10.29  < 2e-16 ***
art          0.5813077  0.0277378   20.96  < 2e-16 ***
```

```
geog                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30355  on 49999  degrees of freedom
Residual deviance: 24122  on 49989  degrees of freedom
AIC: 24144

Number of Fisher Scoring iterations: 6
```

Buying Prediction and Customer Profiling

Based on your logistic model, predict the probability of subscribing for each customer. Then, profile the top decile and the bottom decile. Provide your interpretations and recommendations for promotional targeting.

- Interpretation: The coefficient for this variable is NA, indicating potential collinearity or singularity issues in the model.

- Recommendation: Further investigate the relationship between geography and travel book purchases and online subscription behavior to address any data issues and gain insights for targeted promotions.

Overall, the recommendations focus on leveraging customer behavior, preferences, and purchase patterns to develop targeted promotional strategies aimed at increasing subscription rates for the online book club. Additionally, addressing any data issues and refining the model can provide more accurate insights for future promotional targeting efforts.