1. Hypotheses:
    1. Before performing any analysis, for each of the 9 independent variables, predict if it will have a direct, indirect, or no relationship with the dependent variable (CancelledService). State your reason for each prediction.

For the AccountWeeks, I assume that it has an indirect relationship with the dependent variable. Because I choose the first 24 data from the table and calculate the average of the two groups. The average of the group which canceled the service is 76, and the other is 94.23. The former group is greater than the other by 18.23

For the recent renewal, I assume that it has a direct relationship with the dependent variable. Because I choose the last 25 data from the table and find out that in these 25 people there are only 3 people who recently have not to renew their service and canceled the service.

For the data plane, I assume that it has a direct relationship with the dependent variable. Because I choose 26 data from the table and classify these data into 4 groups: the customer does not have a data plan and cancel the service, the customer does not have a data plan and did not cancel the service, the customer has a data plan and cancel their service, and the customer has a data plan and didn't cancel the service. Through the classification, I find out the customers who do not have a data plan and cancel the service is over 50% in these data.

For the data usage, I assume that it has an indirect relationship with the dependent variable. Because in the table, there are a lot of customers who have 0 data usage and keep their service.

For the number of calls made to the customer service department, I assume that it has an indirect relationship with the dependent variable. I observe that those customers who make calls to the customer service department still keep their service, and only a few customers made some calls and cancel their service.

For the average minutes of calls per month, I assume that is no relationship with canceled service or not. I collect 26 records in the table and calculate the average of two groups. The group of canceled the service is 226.3 and the other is 199.7.

For the average number of calls per month, I assume that is no relationship with the dependent variable. I collect 26 records in the table and calculate the average of two groups. Customers who canceled their service have an average of 99 calls per month. And the others have 96 calls in a month.

For the average monthly charge, I assume that is no relationship with canceled service or not. I collect 52 records in the table and calculate the averages of the two groups are 58.4 and 54.82. They are very close so I predict there is no relationship.

For the highest overage fee in the past 12 months, I assume that is a direct relationship with canceled service. I observe that most customers who canceled their service have a higher fee than those who did not cancel the service.

2. From R:
    1. All code from your R script (Code should be presented single-spaced in a fixed-width font. Adjust the font size so that no lines of code extend to the next line in the document)

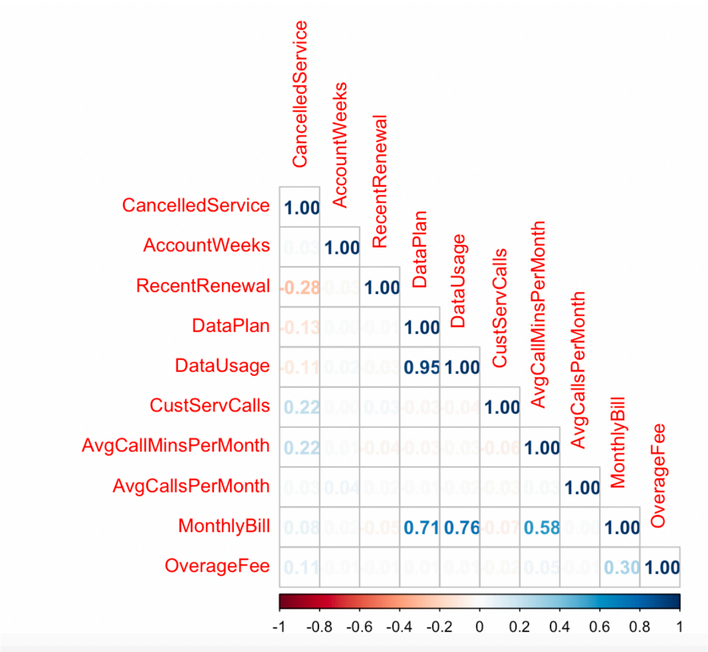
```
# YUEH-TING WU
# MIS 545 Section 02
# Lab06WuY.R
# In this R programming, importing a csv file, displaying some histograms,
# creating a correlation matrix and split the dataset into training and testing.
# And then generating a logistic model.

# Install the tidyverse, corrplot, olsrr, and smotefamily packages
# install.packages("tidyverse")
# install.packages("corrplot")
# install.packages("olsrr")
# install.packages("smotefamily")

# Load the tidyverse, corrplot, olsrr, and smotefamily libraries
library(tidyverse)
library(corrplot)
library(olsrr)
library(smotefamily)

# Set the working directory to Lab06 folder
setwd("~/MIS 545/Lab06")

# Read MobilePhoneSubscribers.csv into a tibble called mobilePhone
mobilePhone <- read_csv(file = "MobilePhoneSubscribers.csv",
                        col_names = TRUE,
                        col_types = "lillnininn")

# Display mobilePhone in the console
print(mobilePhone)

# Display the structure of the mobilePhone in the console
str(mobilePhone)

# Display the summary of mobilePhone in the console
summary(mobilePhone)

# Recreate the displayAllHistograms() function
displayAllHistograms <- function(tibbleDataset) {
  tibbleDataset %>%
    keep(is.numeric) %>%
    gather() %>%
    ggplot() + geom_histogram(mapping = aes(x=value, fill=key),
                              color = "black") +
    facet_wrap (~key, scales = "free") +
    theme_minimal ()
}
```

```
# Call the displayAllHistograms() function, passing into mobilePhone as an
# argument
displayAllHistograms(mobilePhone)

# Display a correlation matrix of mobilePhone
cor(mobilePhone)

# If the tibble has non-numeric values, limit the correlation matrix to
# numeric values to prevent errors
cor(mobilePhone %>%
    keep(is.numeric))

# And rounded to two decimal places
round(cor(mobilePhone), 2)

# Display a correlation plot using the "number" method and limit output to the
# bottom left
corrplot(cor(mobilePhone),
      method = "number",
      type = "lower")

# The correlation plot reveal three pairwise correlations that are above the
# threshold of 0.7. Remove the data plan and data usage variables from the
# tibble
mobilePhone <- select(.data = mobilePhone,
              -DataPlan,
              -DataUsage)

# Spilt data into training and testing
# The set.seed() function is used to ensure that we can get the same result
# every time we run a random sampling process.
set.seed(203)

# Create a vector of 75% randomly sample rows from the orighinal dataset
sampleSet <- sample(nrow(mobilePhone),
                    round(nrow(mobilePhone) * 0.75),
                    replace = FALSE)
# Put the records from the 75% sample into mobilePhoneTraining
mobilePhoneTraining <- mobilePhone[sampleSet, ]
# Put the records from the 25% sample into mobilePhoneTesting
mobilePhoneTesting <- mobilePhone[-sampleSet, ]

# Check if we have a class imbalance issue in CancelledService
summary(mobilePhoneTraining$CancelledService)
```

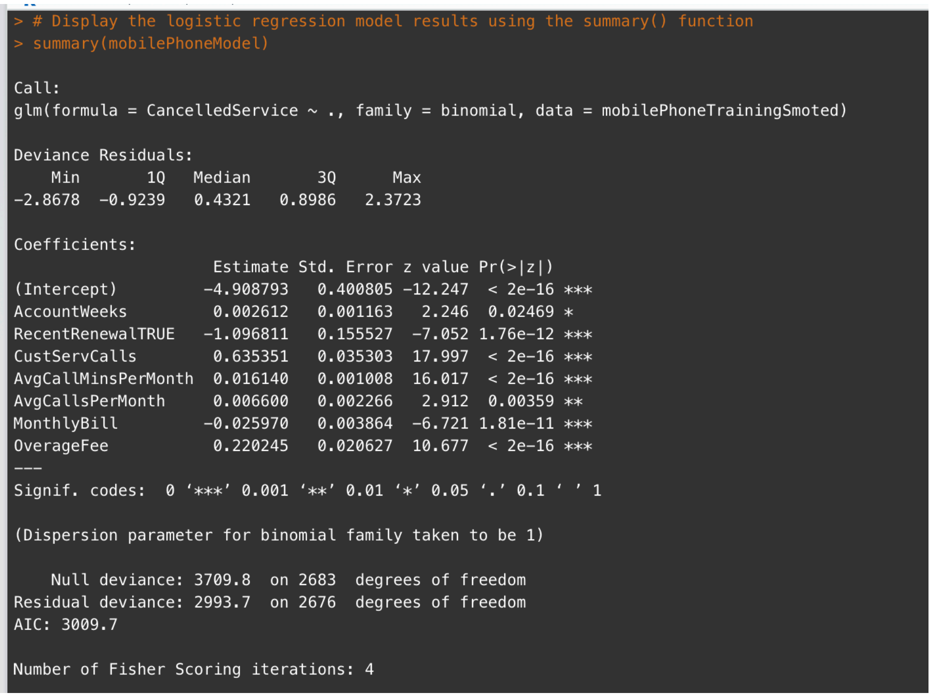
```
# Deal with class imbalance using the SMOTE technique, using a duplicate size
# of 3. And save the result into a new tibble called mobilePhoneTrainingSmoted
mobilePhoneTrainingSmoted <-
  tibble(SMOTE(X = data.frame(mobilePhoneTraining),
               target = mobilePhoneTraining$CancelledService,
               dup_size = 3)$data)

# Convert CancelledService and RecentRenewal back into logical types
mobilePhoneTrainingSmoted <- mobilePhoneTrainingSmoted %>%
  mutate(CancelledService = as.logical(CancelledService),
         RecentRenewal = as.logical(RecentRenewal))

# Get rid of the "class" column in the tibble
mobilePhoneTrainingSmoted <- mobilePhoneTrainingSmoted %>%
  select(-class)

# Check for class imbalance on the smoted dataset
summary(mobilePhoneTrainingSmoted)

# Generate the logistic regression model and save it in an object called
# mobilePhoneModel
mobilePhoneModel <- glm(data = mobilePhoneTrainingSmoted,
                        family = binomial,
                        formula = CancelledService ~ .)

# Display the logistic regression model results using the summary() function
summary(mobilePhoneModel)

# Calculate the odds ratios for each of the 7 independent variable coefficients
exp(coef(mobilePhoneModel)["AccountWeeks"])
exp(coef(mobilePhoneModel)["RecentRenewal"])
exp(coef(mobilePhoneModel)["CustServCalls"])
exp(coef(mobilePhoneModel)["AvgCallMinsPerMonth"])
exp(coef(mobilePhoneModel)["AvgCallsPerMonth"])
exp(coef(mobilePhoneModel)["MonthlyBill"])
exp(coef(mobilePhoneModel)["OverageFee"])

# Use the model to predict outcomes in the testing dataset as described
mobilePhonePrediction <- predict(mobilePhoneModel,
                                 mobilePhoneTesting,
                                 type = "response")

# Treat anything below or equal to 0.5 as a 0, anything above 0.5 as a 1
mobilePhonePrediction <- ifelse(mobilePhonePrediction >= 0.5, 1, 0)

# Generate a confusion matrix of predictions
```

```
mobilePhonePredictionConfustionMatrix <-
  table(mobilePhoneTesting$CancelledService, mobilePhonePrediction)

print(mobilePhonePredictionConfustionMatrix)

# Calculate the false positive rate, 130/(282+130)
# This predict that the customer would cancel service, but they did not
mobilePhonePredictionConfustionMatrix[1, 2] /
  (mobilePhonePredictionConfustionMatrix[1, 1] +
     mobilePhonePredictionConfustionMatrix[1, 2])

# Calculate the false negative rate, 33/(33+93)
# This predict that the customer would not cancel service, but they did
mobilePhonePredictionConfustionMatrix[2, 1] /
  (mobilePhonePredictionConfustionMatrix[2, 1] +
     mobilePhonePredictionConfustionMatrix[2, 2])

# Calculate the prediction accuracy by dividing the number of true positive and
# true negatives by the total amount of predictions in the testing dataset
sum(diag(mobilePhonePredictionConfustionMatrix)) / nrow(mobilePhoneTesting)
```

2. The correlation plot

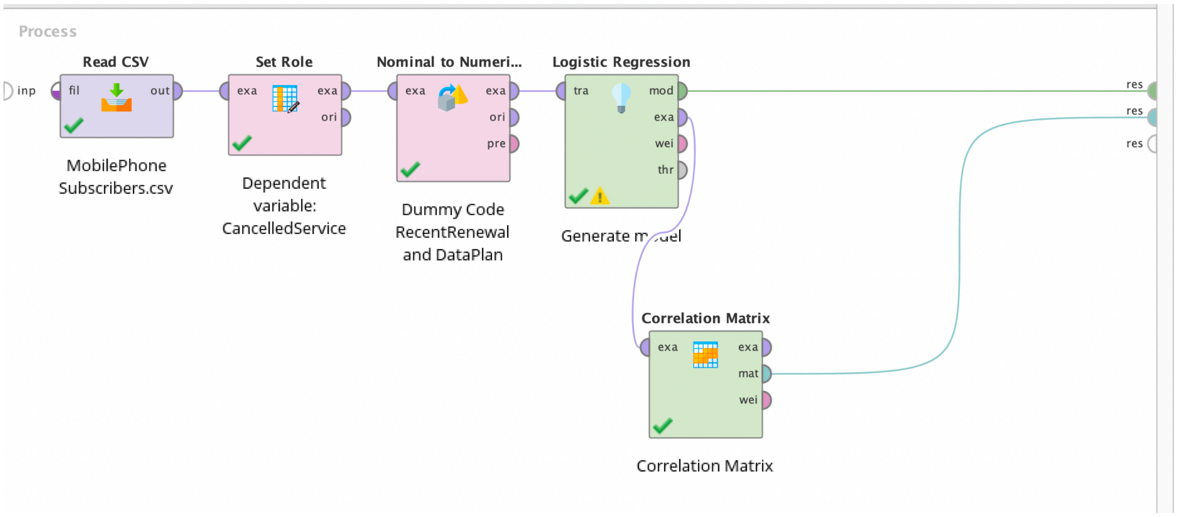


3. The model summary

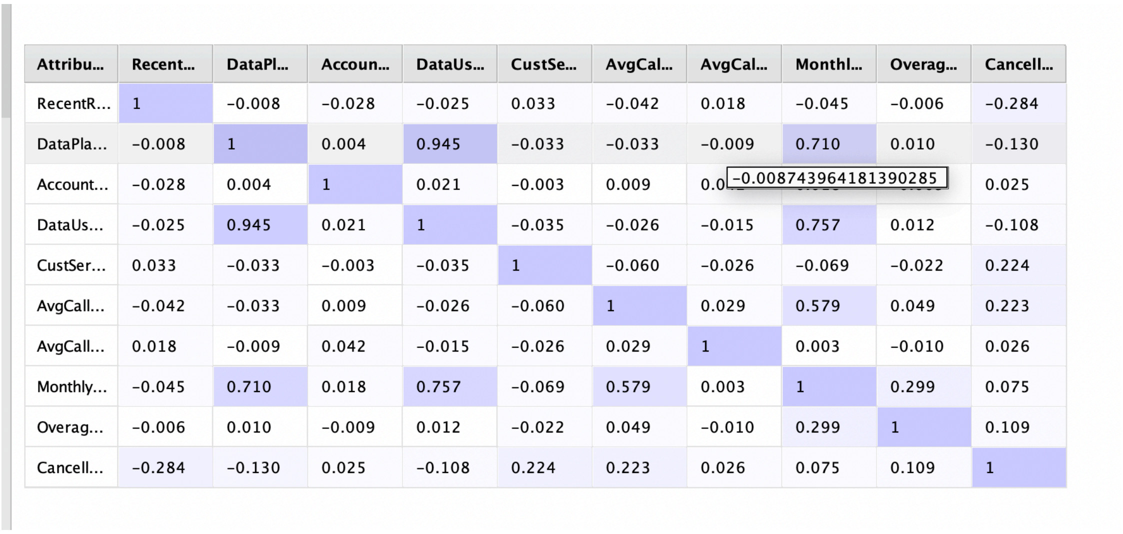


```
> # Display the logistic regression model results using the summary() function
> summary(mobilePhoneModel)

Call:
glm(formula = CancelledService ~ ., family = binomial, data = mobilePhoneTrainingSmoted)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8678  -0.9239   0.4321   0.8986   2.3723

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -4.908793   0.400805 -12.247  < 2e-16 ***
AccountWeeks         0.002612   0.001163   2.246  0.02469 *
RecentRenewalTRUE   -1.096811   0.155527  -7.052 1.76e-12 ***
CustServCalls        0.635351   0.035303  17.997  < 2e-16 ***
AvgCallMinsPerMonth  0.016140   0.001008  16.017  < 2e-16 ***
AvgCallsPerMonth     0.006600   0.002266   2.912  0.00359 **
MonthlyBill         -0.025970   0.003864  -6.721 1.81e-11 ***
OverageFee           0.220245   0.020627  10.677  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3709.8  on 2683  degrees of freedom
Residual deviance: 2993.7  on 2676  degrees of freedom
AIC: 3009.7

Number of Fisher Scoring iterations: 4
```

3. From RapidMiner:
    1. A screenshot of your process

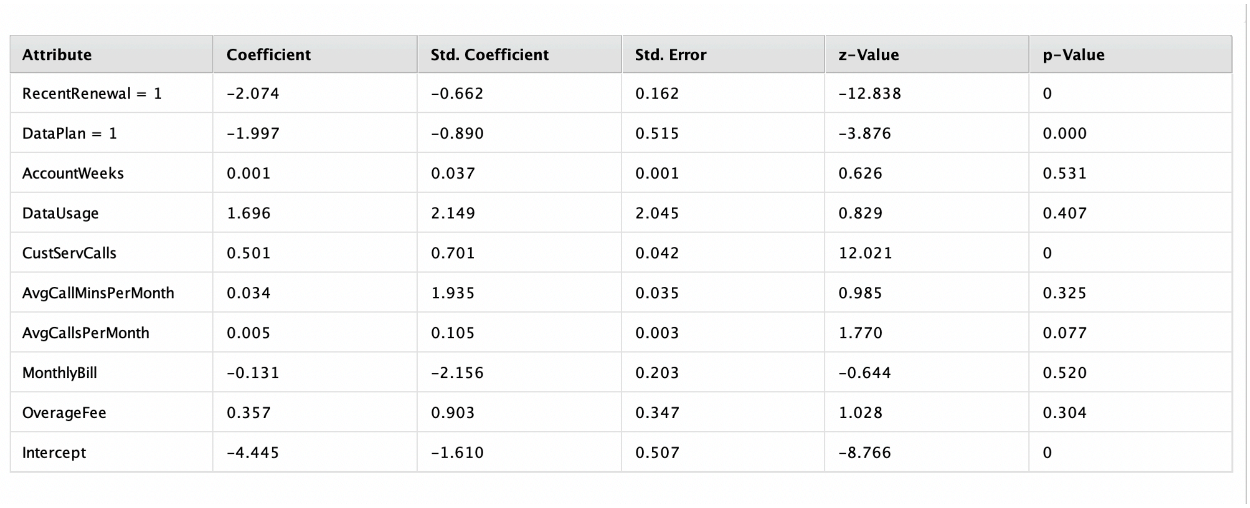


    2. A screenshot of the generated correlation matrix



| Attribu... | Recent... | DataPl... | Accoun... | DataUs... | CustSe... | AvgCal... | AvgCal... | Monthl... | Overag... | Cancell... |
|---|---|---|---|---|---|---|---|---|---|---|
| RecentR... | 1 | -0.008 | -0.028 | -0.025 | 0.033 | -0.042 | 0.018 | -0.045 | -0.006 | -0.284 |
| DataPla... | -0.008 | 1 | 0.004 | 0.945 | -0.033 | -0.033 | -0.009 | 0.710 | 0.010 | -0.130 |
| Account... | -0.028 | 0.004 | 1 | 0.021 | -0.003 | 0.009 | 0.0 | -0.008743964181390285 | | 0.025 |
| DataUs... | -0.025 | 0.945 | 0.021 | 1 | -0.035 | -0.026 | -0.015 | 0.757 | 0.012 | -0.108 |
| CustSer... | 0.033 | -0.033 | -0.003 | -0.035 | 1 | -0.060 | -0.026 | -0.069 | -0.022 | 0.224 |
| AvgCall... | -0.042 | -0.033 | 0.009 | -0.026 | -0.060 | 1 | 0.029 | 0.579 | 0.049 | 0.223 |
| AvgCall... | 0.018 | -0.009 | 0.042 | -0.015 | -0.026 | 0.029 | 1 | 0.003 | -0.010 | 0.026 |
| Monthly... | -0.045 | 0.710 | 0.018 | 0.757 | -0.069 | 0.579 | 0.003 | 1 | 0.299 | 0.075 |
| Overag... | -0.006 | 0.010 | -0.009 | 0.012 | -0.022 | 0.049 | -0.010 | 0.299 | 1 | 0.109 |
| Cancell... | -0.284 | -0.130 | 0.025 | -0.108 | 0.224 | 0.223 | 0.026 | 0.075 | 0.109 | 1 |

3. A screenshot of the logistic regression model results



| Attribute | Coefficient | Std. Coefficient | Std. Error | z-Value | p-Value |
|---|---|---|---|---|---|
| RecentRenewal = 1 | -2.074 | -0.662 | 0.162 | -12.838 | 0 |
| DataPlan = 1 | -1.997 | -0.890 | 0.515 | -3.876 | 0.000 |
| AccountWeeks | 0.001 | 0.037 | 0.001 | 0.626 | 0.531 |
| DataUsage | 1.696 | 2.149 | 2.045 | 0.829 | 0.407 |
| CustServCalls | 0.501 | 0.701 | 0.042 | 12.021 | 0 |
| AvgCallMinsPerMonth | 0.034 | 1.935 | 0.035 | 0.985 | 0.325 |
| AvgCallsPerMonth | 0.005 | 0.105 | 0.003 | 1.770 | 0.077 |
| MonthlyBill | -0.131 | -2.156 | 0.203 | -0.644 | 0.520 |
| OverageFee | 0.357 | 0.903 | 0.347 | 1.028 | 0.304 |
| Intercept | -4.445 | -1.610 | 0.507 | -8.766 | 0 |

4. Answer the following question in a sentence: Which, if any, of your predictions were incorrect. Explain why this might be the case.

There are four incorrect predictions which are CusServCalls, AvgCallMinsPerMonth, AvgCallsPerMonth, and MonthlyBill. In each of my predictions, I just consider one independent variable at a time and did not consider other independent variables that also have a correlation with others.

5. Answer the following question in a sentence: Why is DataPlan highly correlated with DataUsage?

I think it is because when we joined in telecommunications, the data usage is included in the data plan. That’s why DataPlan is highly correlated with DataUsage.

6. Answer the following question in a sentence: Why is MonthlyBill highly correlated with DataPlan and DataUsage?

II think that is because the monthly bill depends on which data plan the customer used and how much data usage does the customer use in a month.