Hypotheses:

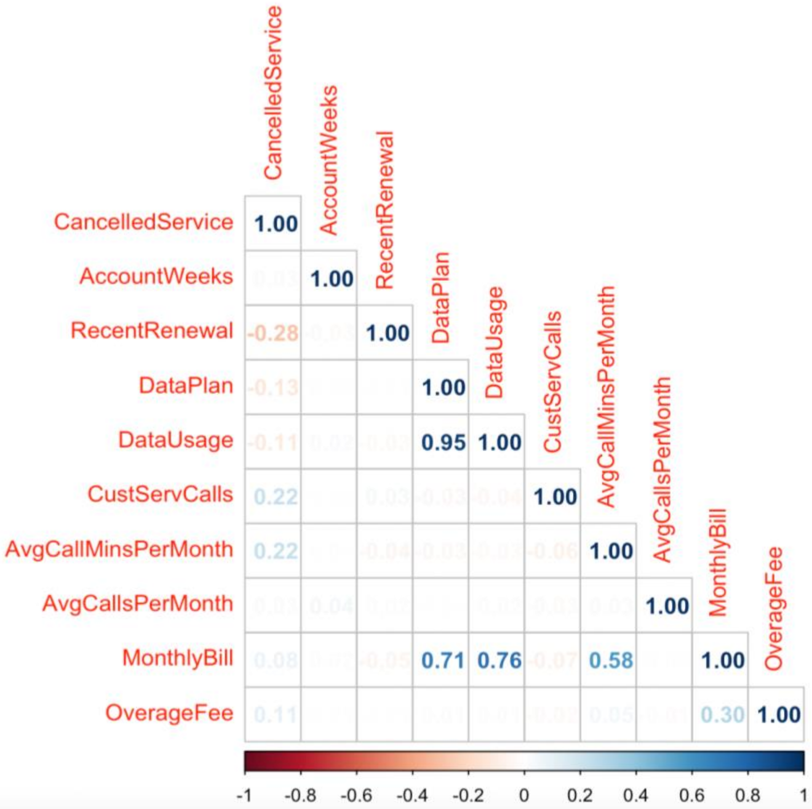
1. **AccountWeeks** – This should probably have no relationship as the customer could cancel the service any time even after using it for weeks together
2. **Recent Renewal** – This should have a direct relationship in the negative direction as, if the customer has renewed the service recently, he is most likely not going to cancel.
3. **Data Plan** – This should have a direct relationship in the negative direction as, if the customer has an active plan, he most likely won't cancel the service.
4. **Data Usage** – This should have an indirect relationship in the negative direction as, if the customer has used enough data, it means he mostly likes the service.
5. **CustServCalls** – This could probably be an indirect relationship in the positive direction as, if the customer is making a lot of calls to the customer care center, then he is probably facing a lot of issues with the service.
6. **AvgCallMinsPerMonth** – This could probably have an indirect relationship in the negative direction as if the customer is making a lot of calls, then he is probably happy and not going to cancel the service.
7. **MonthlyBill** – This could probably have no relationship at all as, if the bill is too much and the customer isn't happy with the service, he could cancel it or if it is cheap enough, then he could continue the service.
8. **Overage fee** – This will probably have no relationship at all as, he could be paying the overage fee by his own choice.

R script:

```
# Aadhithya Dinesh
# MIS 545 Section 02
# Lab06DineshA.R
# Import a dataset of mobile phone plan subscribers and generate a multiple
# logistic regression model that will predict if a customer will
# cancel their contract based on a number of different factors.

#install.packages ("tidyverse")
#library (tidyverse)
#install.packages ('corrplot')
#library (corrplot)
#install.packages("olsrr")
#library (olsrr)
#install.packages("smotefamily")
#library (smotefamily)

# Set the working directory
setwd("~/MIS/Classes/MIS545/Assignments/Lab06")

mobilePhone <- read_csv(file = "MobilePhoneSubscribers.csv",
               col_types = "lillnininn",
                       col_names = TRUE)


# print the mobilePhone tibble
print(mobilePhone)

#print the structure of mobilePhone
str(mobilePhone)

#print the summary of mobilePhone
print(summary(mobilePhone))

# define the function to display all histograms
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 function
displayAllHistograms(mobilePhone)

# rounding the correlation to 2 decimal places
print(round(cor(mobilePhone  %>% keep(is.numeric)),2))

# displaying the correlation plot using number method
corrplot(cor(mobilePhone),
         method = "number",
         type = "lower")

mobilePhone <- select(mobilePhone, -c(DataPlan, DataUsage))

# random dataset with 203 as seed
set.seed(203)

# creating a vector of 75% ramdomly sampled rows
sampleSet<- sample (nrow(mobilePhone),
                    round(nrow(mobilePhone)*0.75),
                    replace= FALSE)
```

```
# assign 75% to mobilePhoneTraining
mobilePhoneTraining<- mobilePhone[sampleSet, ]

# assign 25% to mobilePhoneTesting
mobilePhoneTesting<- mobilePhone[-sampleSet, ]

# checking imbalance
print(summary(mobilePhoneTraining$CancelledService))

# storing the magnitude of imbalance
classImbalanceMagnitude<- 1256/357

# dealing with class imbalance using SMOTE technique
mobilePhoneTrainingSmoted<-
  tibble(SMOTE(X=mobilePhoneTraining,
               target=mobilePhoneTraining$CancelledService,
               dup_size = 3)$ data)

print(summary(mobilePhoneTrainingSmoted))

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

# deleting "class" column
mobilePhoneTrainingSmoted<- mobilePhoneTrainingSmoted %>%
  select(-class)

summary(mobilePhoneTrainingSmoted)

# generating the logistic regression model
mobilePhoneModel<- glm(data= mobilePhoneTrainingSmoted,
                       family=binomial,
                       formula= CancelledService ~ . )

# displaying the logistic regression model
summary(mobilePhoneModel)

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

# predicted outcomes
mobilePhonePrediction<- predict(mobilePhoneModel,
                                mobilePhoneTesting,
                                type="response")
# display mobilePhonePrediction
print(mobilePhonePrediction)

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

# displaying mobilePhonePrediction
print(mobilePhonePrediction)

# creating confusion matrix
mobilePhoneConfusionMatrix<- table(mobilePhoneTesting$CancelledService,
                                   mobilePhonePrediction)

# displaying mobilePhoneConfusionMatrix
print(mobilePhoneConfusionMatrix)

# calculating false positive
mobilePhoneConfusionMatrix[1,2]/
  (mobilePhoneConfusionMatrix[1,2]+
    mobilePhoneConfusionMatrix[1,1])

# calculating false negative
mobilePhoneConfusionMatrix[2,1]/
  (mobilePhoneConfusionMatrix[2,1]+
    mobilePhoneConfusionMatrix[2,2])

# Calculating model prediction accuracy
sum(diag(mobilePhoneConfusionMatrix))/ nrow(mobilePhoneTesting)
```

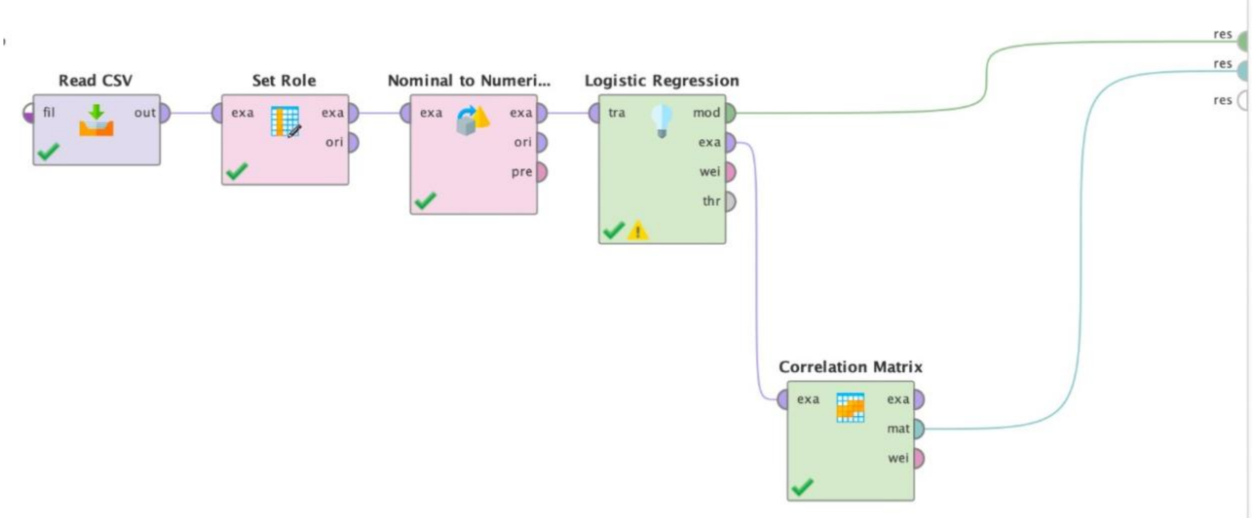
Correlation Plot:



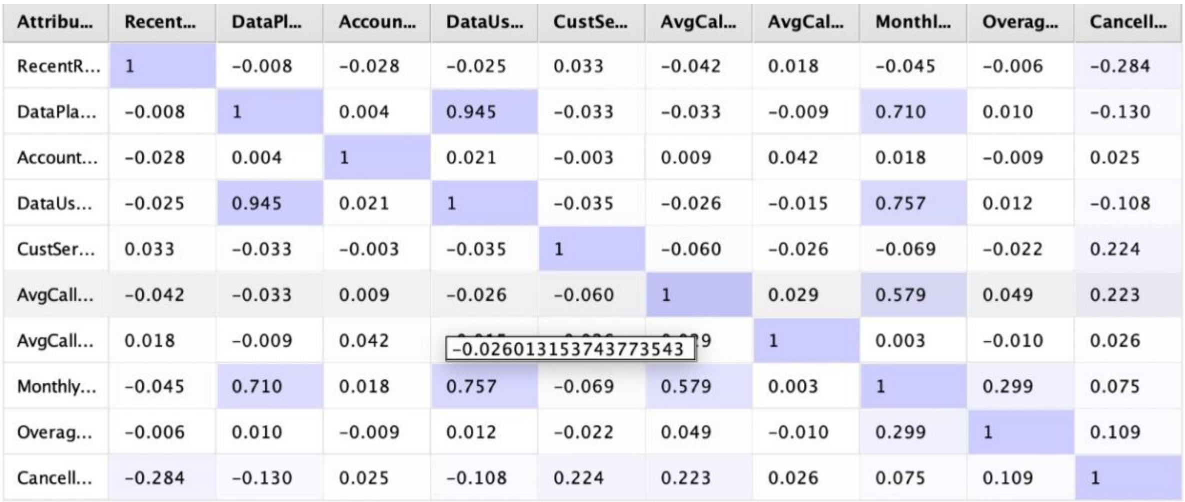
Model Summary:

```
> 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
```

Rapid Miner Process:

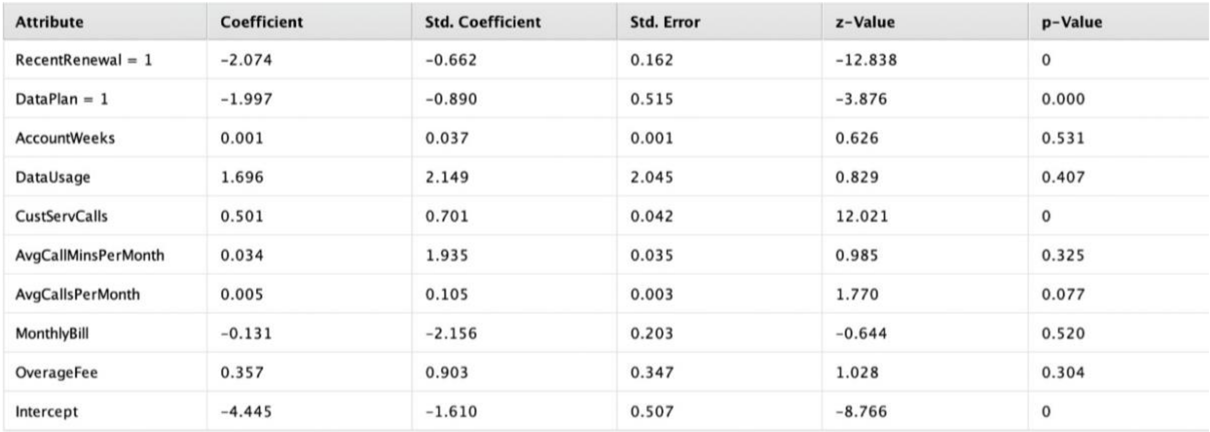


Correlation Matrix:



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	0.042	0.018	-0.009	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	-0.026013153743773543	[illegible]	9	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

Model Summary:



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

Answers:

1. My hypotheses were right for most of the independent variables.
2. Data Plan and Data usage have a correlation of 0.95, the reason being customer can have data usage only if he has a data plan.
3. Monthly Bill and Data Plan have a correlation of 0.71, Monthly Bill and Data Usage have a correlation of 0.76. The reason being customer will tend to pay a monthly bill higher if he has a data plan and uses the data.