Name: Praful Patil, Sanjana Belde, Sreejanya Onteddu, Tulasi Smrithi Yaralagadda
NetID: PVP220001, SXB220133, SXO220000, TXY220003

Group No: 13 Assignment 5

PART 1:

a) a)

```
# 1).
# A).
# a)
# Read in the CSV file
data <- read.csv("cell2cell.csv")

# Get the total number of customers (rows)
total_customers <- nrow(data)
total_customers
```

b) Number of customers in calibration and validation set

```
# b)
# Count how many are in the calibration set
calibration_count <- sum(data$calibrat == 1)

# Count how many are in the validation set
validation_count <- sum(data$calibrat == 0)

calibration_count
validation_count
```

c)

```
# c) Churn rate in calibration set
calibration_churn_rate <- mean(data$churn[data$calibrat == 1] == 1)
```

d)

```
# d) Churn rate in validation set
validation_churn_rate <- mean(data$churn[data$calibrat == 0] == 1)
```

Results for a), b) c) and d)

```
> # Results
> cat("Total number of customers:", total_customers, "\n")
Total number of customers: 69626
> cat("Customers in calibration set:", calibration_count, "\n")
Customers in calibration set: 39186
> cat("Customers in validation set:", validation_count, "\n")
Customers in validation set: 30440
> cat("Churn rate in calibration set:", format(calibration_churn_rate, digits=6), "\n")
Churn rate in calibration set: 0.5
> cat("Churn rate in validation set:", format(validation_churn_rate, digits=6), "\n")
Churn rate in validation set: 0.0195466
```

Name: Praful Patil, Sanjana Belde, Sreejanya Onteddu, Tulasi Smrithi Yaralagadda
NetID: PVP220001, SXB220133, SXO220000, TXY220003

Group No: 13 Assignment 5

1)

b) Data Cleaning

a)

```
# Selecting relevant columns
Data <- data %>%
  select(churndep, revenue:retcall, calibrat)
```

b)

```
# b).
# Split the data into training and validation sets
training_set <- Data %>%
  filter(calibrat == 1) %>%
  select(-calibrat)

validation_set <- Data %>%
  filter(calibrat == 0) %>%
  select(-calibrat)
```

c)

Running logistic model

```
# C) Running logistic model for training dataset
logistic_model <- glm(churndep ~ ., data = training_set, family = binomial)

# Summarizing the model to check coefficients and p-values
summary_logistic <- summary(logistic_model)

# P value and Odds ratio of Top and Bottom two variables
sorted_odds_desc <- sort(exp(coef(logistic_model)), decreasing = TRUE)
top_bottom_odds <- c(head(sorted_odds_desc, 2), tail(sorted_odds_desc, 2))
top_bottom_pvalues <- summary_logistic$coefficients[c(names(head(sorted_odds_desc, 2)), names(tail(sorted_odds_desc, 2))), "Pr(>|z|)"]
top_bottom_odds
top_bottom_pvalues
```

```
> summary_logistic

Call:
glm(formula = churndep ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4750  -1.1379  -0.6788   1.1472   2.5022

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.432e-01  9.625e-02   1.488 0.136707
revenue      1.839e-03  8.070e-04   2.279 0.022693 *
mou         -2.846e-04  5.022e-05  -5.667 1.46e-08 ***
recchrge    -3.146e-03  8.980e-04  -3.503 0.000460 ***
directas     1.247e-03  5.991e-03   0.208 0.835171
overage      8.270e-04  2.840e-04   2.912 0.003589 **
```

Odds Ratios for all variables in Data under training set

Name: Praful Patil, Sanjana Belde, Sreejanya Onteddu, Tulasi Smrithi Yaralagadda
NetID: PVP220001, SXB220133, SXO220000, TXY220003

Group No: 13 Assignment 5

```
> sorted_odds_desc
    retcall      refurb     occhmkr    uniqsubs      mcycle (Intercept)     occstud     marryun    children    prizmrur    marryyes
  2.2889200   1.2604818   1.2353042   1.2053592   1.1559566   1.1540088   1.1335067   1.1222327   1.0925166   1.0712510   1.0610166
     phones     occcler    prizmtwn       pcown       truck    creditcd      models     dropvce        roam     blckvce     changer
  1.0488884   1.0472055   1.0470033   1.0357095   1.0343554   1.0339514   1.0169642   1.0119703   1.0073437   1.0065287   1.0022736
   callwait     revenue     eqpdays    directas          rv    outcalls    unansvce     overage      setprc      mourec    opeakvce
  1.0020989   1.0018404   1.0014687   1.0012473   1.0011574   1.0010584   1.0009444   1.0008273   1.0006528   1.0001350   0.9998118
    ownrent         mou     changem     peakvce        age2     incalls    recchrge     dropblk        age1    custcare    callfwdv
  0.9997717   0.9997155   0.9995122   0.9993158   0.9988733   0.9968604   0.9968593   0.9967456   0.9966805   0.9936072   0.9925639
   newcelln      travel      income     occprof     mailord      months    retcalls     occcrft      occret    threeway     prizmub
  0.9912671   0.9906371   0.9873534   0.9867392   0.9817660   0.9786853   0.9783057   0.9697221   0.9686486   0.9662418   0.9651080
   mailflag     occself       refer    newcelly     incmiss     setprcm     mailres    retaccpt     web4glte     credita    actvsubs
  0.9616924   0.9508921   0.9473410   0.9331411   0.9088304   0.9050376   0.8945173   0.8701953   0.8586624   0.8372925   0.8129321
   creditaa
  0.6977511
```

```
> top_bottom_odds
  retcall    refurb  actvsubs  creditaa
2.2889200 1.2604818 0.8129321 0.6977511
> top_bottom_pvalues
     retcall       refurb     actvsubs     creditaa
2.199318e-05 7.761796e-13 2.272651e-13 7.239558e-25
```

Two highest Odds ratios

retcall (Customer has made a call to retention team): With an odds ratio of 2.288920, customers who have made a call to the retention team are 2.29 times more likely to churn than those who have not, holding other variables constant. This is statistically significant (p = 2.199318e-05).

refurb (Handset is refurbished): An odds ratio of 1.2604818 means that customers with a refurbished handset are 1.26 times more likely to churn than those with a new handset, ceteris paribus. This relationship is also statistically significant (p = 7.761796e-13).

Two lowest Odds ratios

actvsubs (Number of Active Subs): As a count variable, the odds ratio of 0.8129321 suggests that for each additional active subscription, the odds of churning decrease by approximately 18.7%. More active subscriptions are associated with a lower probability of churn. The significance of this relationship is confirmed by the p-value (p = 2.272651e-13).

creditaa (High credit rating - aa): Since this is a dummy variable, the odds ratio of 0.6977511 indicates that customers with a high credit rating are approximately 30.2% less likely to churn than those without a high credit rating, all other factors being equal. The statistical significance is very strong (p = 7.239558e-25).

D) Predicting attrition probabilities

```
# D)
# Use the model to predict attrition probabilities
validation_set$attrition_probability <- predict(logistic_model, newdata = validation_set, type = "response")

# Arrange the data in descending order of attrition probabilities
validation_set <- validation_set %>%
  arrange(desc(attrition_probability))

# Display the top 5 attrition probabilities
head(validation_set$attrition_probability)
```

```
> head(validation_set$attrition_probability)
[1] 0.9998094 0.9976956 0.9962896 0.9945006 0.9704669 0.9549060
```

Name: Praful Patil, Sanjana Belde, Sreejanya Onteddu, Tulasi Smrithi Yaralagadda
NetID: PVP220001, SXB220133, SXO220000, TXY220003

Group No: 13 Assignment 5

PART 2:

a) Data frame (i.e., data matrix) with odds ratios and p-values as 2 columns

```
# A)
# a)
OR <- sort(exp(coef(logistic_model)), decreasing = TRUE)

# b)
pvales <- coef(summary(logistic_model))[,'Pr(>|z|)']

# c)
df1 <- data.frame(OR, pvales)
```

Output:

Filter

	OR	pvales
retcall	2.2889200	1.367073e-01
refurb	1.2604818	2.269327e-02
occhmkr	1.2353042	1.456866e-08
uniqsubs	1.2053592	4.601099e-04
mcycle	1.1559566	8.351709e-01
(Intercept)	1.1540088	3.589164e-03
occstud	1.1335067	4.890681e-04
marryun	1.1222327	1.755390e-19
children	1.0925166	9.762995e-10
prizmrur	1.0712510	1.024259e-01

b) Function to check non-missing values

Function:

```
# 2)
# b) -> a), b)
# Function to calculate standard deviation if there are non-missing values
calculate_sd <- function(x) {
  # Check if there are any non-missing values
  if (sum(!is.na(x)) > 0) {
    # Calculate and return standard deviation, excluding NA values
    return(sd(x, na.rm = TRUE))
  } else {
    # Return NA if all values are missing
    return(NA)
  }
}
```

Name: Praful Patil, Sanjana Belde, Sreejanya Onteddu, Tulasi Smrithi Yaralagadda
NetID: PVP220001, SXB220133, SXO220000, TXY220003

Group No: 13 Assignment 5

Apply the custom function (#calculate_sd) to each column in the dataframe to calibrate data i.e. Validation Set

```
# Apply the custom function (#calculate_sd) to each column in the dataframe to calibr
standard_deviations <- sapply(validation_set, calculate_sd)

# b) c)
# The result is a named vector of standard deviations
df2 <- data.frame(standard_deviations)
```

standard_deviations	VarName
24.12525955	recchrge
2.35230712	directas
93.80650335	overage
10.03972894	roam
249.20432818	changem
37.55775941	changer

c) Merging df1 and df2

```
# c) a)
# Add a column with row names to df1
df1$VarName <- row.names(df1)

# Add a column with row names to df2
df2$VarName <- row.names(df2)

# c) b)
# Merge the data frames based on VarName
# all = FALSE ensures that only the matched rows are kept
df_merged <- merge(df1, df2, by = "VarName", all = FALSE)
```

d) e) Rounding the decimals and then exporting the results to csv file.

```
# d)
# Round numeric columns to 5 decimal points
df_merged[sapply(df_merged, is.numeric)] <- round(df_merged[sapply(df_merged, is.numeric)], 5)


df_sorted <- df_merged %>%
      filter(pvales < 0.05) %>%
      arrange(-OR)

# e)
# Export to CSV
write.csv(df_sorted, "filtered_data.csv", row.names = FALSE)
```

Name: Praful Patil, Sanjana Belde, Sreejanya Onteddu, Tulasi Smrithi Yaralagadda
NetID: PVP220001, SXB220133, SXO220000, TXY220003

Group No: 13 Assignment 5

CSV file written from R

filtered_data

VarName	OR	pvales	standard_deviations
refurb	1.26048	0.02269	0.33938
occhmkr	1.2353	0	0.05665
uniqsubs	1.20536	0.00046	0.83754
occstud	1.13351	0.00049	0.08808
marryun	1.12223	0	0.48549
children	1.09252	0	0.4269
phones	1.04889	0.03682	1.35202
occcler	1.04721	0.01351	0.1398
prizmtwn	1.047	0.00298	0.35425
creditcd	1.03395	0.00324	0.46754
models	1.01696	0.00204	0.92421
callwait	1.0021	0	5.76557
revenue	1.00184	0	44.37081

f)

Calculating economic importance for each variable in CSV file. Following are the steps followed to calculate importance.

- Opened the CSV file in Excel
- Checked if the variable is Dummy or Not (performed manually by using Min and Max values from Cell2Cell Data Documentation.xls) and created a column adjacent to variable name
- Calculated X value for dummy variables : OR^SD, for non-dummy variables : OR
- Importance: For X>1, importance = X; For X<1, importance = 1/X.

VarName	OR	pvales	standard_deviations	Dummy/not	X	Economic Importance
eqpdays	1.00147	0	250.17919	Non-Dummy	1.00147	1.444110221
creditaa	0.69775	2.00E-05	0.3489	Dummy	0.881996529	1.433178072
refurb	1.26048	0.02269	0.33938	Dummy	1.081732539	1.26048
occhmkr	1.2353	0	0.05665	Dummy	1.012042868	1.2353
credita	0.83729	0.01485	0.31078	Dummy	0.94630552	1.194329324
uniqsubs	1.20536	0.00046	0.83754	Non-Dummy	1.20536	1.169333909
retaccpt	0.8702	0.03685	0.12848	Dummy	0.982295737	1.149161112
actvsubs	0.81293	0.02223	0.62519	Non-Dummy	0.81293	1.138240071
occstud	1.13351	0.00049	0.08808	Dummy	1.011099243	1.13351
marryun	1.12223	0	0.48549	Dummy	1.057582484	1.12223
children	1.09252	0	0.4269	Dummy	1.038497629	1.09252
revenue	1.00184	0	44.37081	Non-Dummy	1.00184	1.084986204
peakvce	0.99932	0	107.34461	Non-Dummy	0.99932	1.075751153
newcelly	0.93314	0.01218	0.39458	Dummy	0.97306446	1.071650556
phones	1.04889	0.03682	1.35202	Non-Dummy	1.04889	1.06666317
occcler	1.04721	0.01351	0.1398	Dummy	1.006469741	1.04721
prizmtwn	1.047	0.00298	0.35425	Dummy	1.016403407	1.047
outcalls	1.00106	0	35.6397	Non-Dummy	1.00106	1.038479967
unansvce	1.00094	0.00016	39.79398	Non-Dummy	1.00094	1.038096524
creditcd	1.03395	0.00324	0.46754	Dummy	1.015731951	1.03395
age2	0.99887	6.00E-05	24.05722	Non-Dummy	0.99887	1.027573327
mourec	1.00014	0	167.84632	Non-Dummy	1.00014	1.023775066
mailord	0.98177	8.00E-04	0.48227	Dummy	0.991166341	1.018568504
opeakvce	0.99981	0	94.80353	Non-Dummy	0.99981	1.01817762
models	1.01696	0.00204	0.92421	Non-Dummy	1.01696	1.015664588
callwait	1.0021	0	5.76557	Non-Dummy	1.0021	1.012168442
directas	1.00125	0.00958	2.35231	Non-Dummy	1.00125	1.002942873
setprc	1.00065	0.00189	57.65466	Dummy	1.038173952	1.00065

Name: Praful Patil, Sanjana Belde, Sreejanya Onteddu, Tulasi Smrithi Yaralagadda
NetID: PVP220001, SXB220133, SXO220000, TXY220003

Group No: 13 Assignment 5

PART 3

a.

Variable Name	Variable Description	Dummy/not	OR	P-Value	Standard Deviations	Importance	Actionable
eqpdays	Number of days of the current equipment	Non-Dummy	1.00147	0	250.179	1.44411	Yes
creditaa	High credit rating - aa	Dummy	0.69775	0.00002	0.34890	1.433178	Yes
refurb	Handset is refurbished	Dummy	1.26048	0.02269	0.33938	1.260480	Yes
occhmkr	Occupation - homemaker	Dummy	1.235300	0.00000	0.05665	1.235300	No
actvsubs	Number of Active Subs	Dummy	0.81293	0.02223	0.62519	1.230118	No
uniqsubs	Number of Uniq Subs	Non-Dummy	1.20536	0.00046	0.83754	1.205360	No
credita	Highest credit rating - a	Dummy	0.83729	0.01485	0.31078	1.194329	Yes
retaccpt	Number of previous retention offers accepted	Dummy/not	0.87020	0.03685	0.12848	1.149161	Yes

b. For each actionable and statistically significant predictor variable, the retention action suggested can be the following –

1. Creditaa –
   Negative and actionable
   To increase customer loyalty and retention, provide special discounts and rewards to clients with an excellent credit rating (aa).

2. Refurb –
   Positive and actionable
   To encourage customers to stay on board with their subscription, give customised incentives, such as unique device offers or discounted plans, to those who purchase refurbished handsets.

3. Credita –
   Negative and actionable
   To strengthen customer commitment to Cell2Cell, tailor programs and provide additional benefits to those with the best credit rating (a).

Group No: 13 Assignment 5

4. Retaccpt –
   Positive and actionable
   Create personalized retention offers with extra features and advantages to retain clients who have already accepted prior offers.

5. Eqpdays-
   Positive and actionable
   Its Importance is higher the equipment is older, customers might be experiencing issues or dissatisfaction leading to churn. Offering upgrades or replacements could reduce churn.

c. For each no actionable and statistically significant predictor variable, the information obtained can used as follows-

1. Occhmkr –
   Positive and Non-actionable
   This information can be provided to the marketing team so that they can get a better understanding about their customer base. With the help of this information, they can guide marketing strategies and communications targeted at homemakers.

2. Actvsubs –
   Negative and Non-actionable
   This information can help the Customer Support Teams in knowing how many Subscriptions are active. This knowledge can improve support services and client relations.

3. Uniqsubs –
   Negative and Non-actionable
   This data can be share with the product development teams. They can customize subscription plan and enhancement accordingly.