## Assignment 4 – Model Answers

### Problem 1

Although all the variables are technically numeric, only 'tenure' and 'monthly.charges' are actually continuous. We can run t-tests on these variables. The remaining categorical variables would require chi-squared tests, which you are not required to run in this assignment.

```
t.test_tenure <- t.test(tenure~churn,data=telecom_train)
t.test_tenure

	Welch Two Sample t-test

data:  tenure by churn
t = 24.737, df = 2043, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 18.04073 21.14748
sample estimates:
mean in group 0 mean in group 1
       37.54588        17.95177
```

```
t.test_monthly.charges<- t.test(monthly.charges~churn,data=telecom_train)
t.test_monthly.charges

	Welch Two Sample t-test

data:  monthly.charges by churn
t = -13.818, df = 2097.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.81551 -11.88439
sample estimates:
mean in group 0 mean in group 1
       61.25064        75.10059
```

Both tests are statistically significant, and are likely to be associated with churn.

### Problem 2

We can recycle the code from Assignment 3 to produce a correlation matrix.

In terms of our dependent variable, it seems that paperless billing and monthly charges seem to be slightly associated with churn. Tenure is negatively correlated with churn.

In terms of our explanatory (independent) variables, 'tenure' and 'contract.two.year' seem to be positively correlated, as are monthly.charges with 'streaming.tv' and 'streaming.movies'.

## Problem 3 (3 points)

```
logit.model <- glm(churn ~.,
                   family = binomial, data = telecom_train)

summary(logit.model)
```

```
Call:
glm(formula = churn ~ . - internet - total.charges, family = binomial,
    data = telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7672  -0.6644  -0.2880   0.7344   3.2108

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.079732   0.181657  -5.944 2.78e-09 ***
female             0.032552   0.091327   0.356  0.72152
SeniorCitizen      0.115150   0.118424   0.972  0.33087
partner           -0.051888   0.109326  -0.475  0.63506
dependents        -0.163628   0.125021  -1.309  0.19060
tenure            -0.035086   0.003315 -10.583  < 2e-16 ***
phone.service     -1.245055   0.188513  -6.605 3.99e-11 ***
multiple.lines    -0.064116   0.114852  -0.558  0.57667
online.security   -0.559968   0.117672  -4.759 1.95e-06 ***
online.backup     -0.237403   0.109222  -2.174  0.02974 *
device.protection -0.091455   0.112696  -0.812  0.41707
tech.support      -0.582976   0.118111  -4.936 7.98e-07 ***
streaming.tv      -0.130957   0.120892  -1.083  0.27870
streaming.movies  -0.076039   0.120731  -0.630  0.52881
contract.one.year -0.649845   0.149170  -4.356 1.32e-05 ***
contract.two.year -1.530186   0.257859  -5.934 2.95e-09 ***
paperless.billing  0.272464   0.103742   2.626  0.00863 **
monthly.charges    0.040115   0.003418  11.738  < 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: 4068.6  on 3515  degrees of freedom
Residual deviance: 2944.1  on 3498  degrees of freedom
AIC: 2980.1

Number of Fisher Scoring iterations: 6
```

```
tidy(exp(logit.model$coefficients))
```

```
   names                  x
   <chr>              <dbl>
 1 (Intercept)        0.340
 2 female             1.03
 3 SeniorCitizen      1.12
 4 partner            0.949
 5 dependents         0.849
 6 tenure             0.966
 7 phone.service      0.288
 8 multiple.lines     0.938
 9 online.security    0.571
10 online.backup      0.789
11 device.protection  0.913
12 tech.support       0.558
13 streaming.tv       0.877
14 streaming.movies   0.927
15 contract.one.year  0.522
16 contract.two.year  0.216
17 paperless.billing  1.31
18 monthly.charges    1.04
```

(statistically significant values bolded)

Of the statistically significant variables, all but paperless billing and monthly charges decrease the likelihood of a customer churning. In terms of effect size, the largest positive effect on retention comes from locking a customer in to a 2-year contract (100% * (0,216-1)), which decreases probability of churn by 78,4%. Interestingly, customers who subscribe to paperless billing are 100% * (1,31-1) = 31% more likely to churn than those who do not. Either this service is not implemented in the most optimal way, thus driving customer satisfaction down, or there is omitted variable bias and it is something else about customers who prefer paperless billing that increases their likelihood of churning (perhaps customers who opt for paperless billing are more tech savvy, and thus have a lower threshold for switching telecom service providers).

**Problem 4 (2 points)**

We can obtain the predictions and the confusion matrix by running the following code:

```
predictions_test<-predict(logit.model,newdata=telecom_test,type="response")
pred_cutoff_50_test <- ifelse(predictions_test > 0.5, 1, 0)
table(pred_cutoff_50_test, telecom_test$churn)
```

```
pred_cutoff_50_test    0    1
                  0 2329  430
                  1  251  506
```

Accuracy - our model is able to correctly predict roughly 80% of cases in the dataset: (2329+506)/(2329+506+251+430) = 0.806314

Precision – our model's churn predictions were correct 67% of the time: 506/(506+251) = 0.668428

Recall – our model was able to correctly identify 54% of customers who churned: 506/(506+430) = 0.5405983.

Overall, our model's predictions seem pretty good – it got 81% of all customers in the 'test' dataset correct (accuracy). However, only 67% of its predictions that a customer would churn were correct (precision). More concerning is that it was only able to correctly identify 54% of those customers who actually did churn (recall).

**Problem 5 (2 points)**

From a diagnostic standpoint, the company should look to lock customers in to long-term contracts. Not only are customers on 2-year contract less likely to churn, but the longer a person has been a customer of the service provider, the less likely they are to churn.

The size of the bill is also a factor that affects churning. For example, if a customer is identified as likely to churn, the company could offer a discount on the bill in exchange for signing up for a longer-term contract (e.g., "receive a 20% discount when you sign up for 2 years").