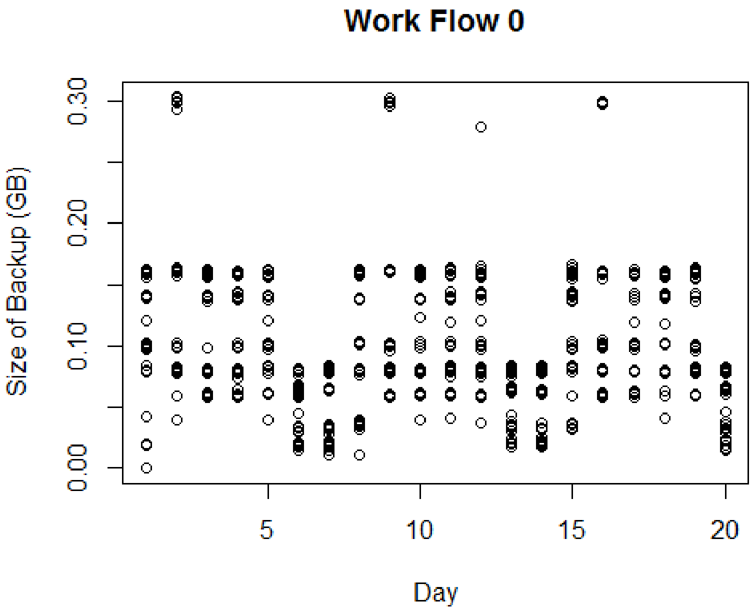
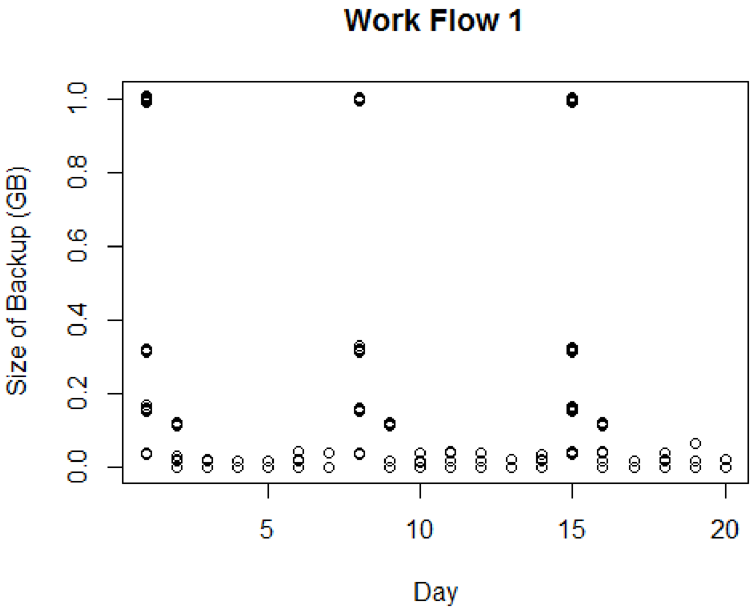
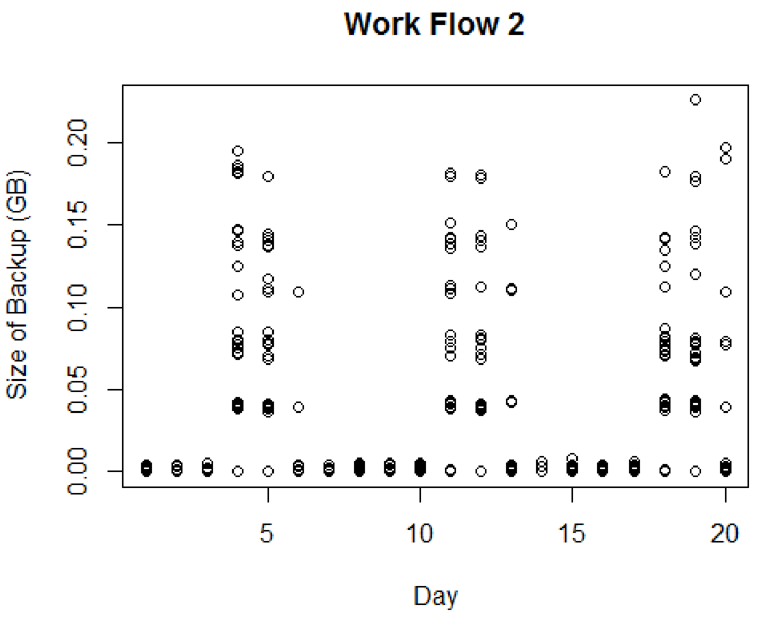
Kevin Wang (604036846)

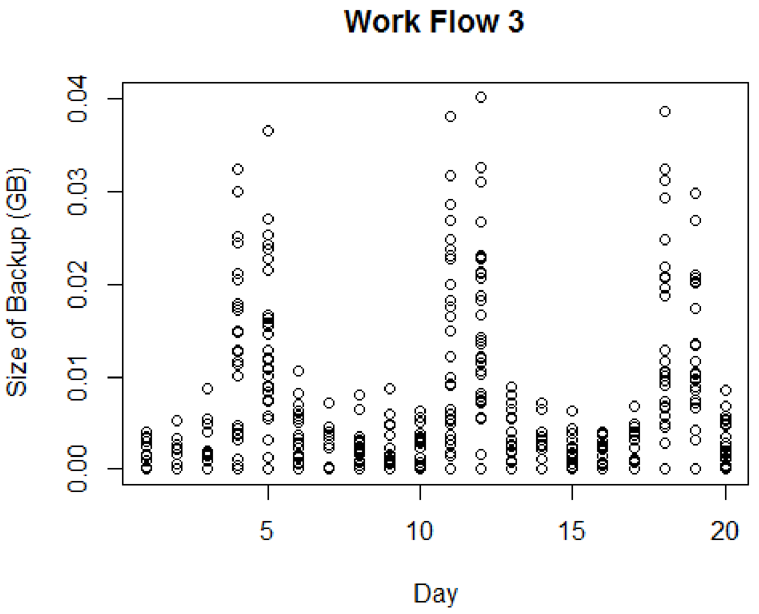
EE 239AS Project 1

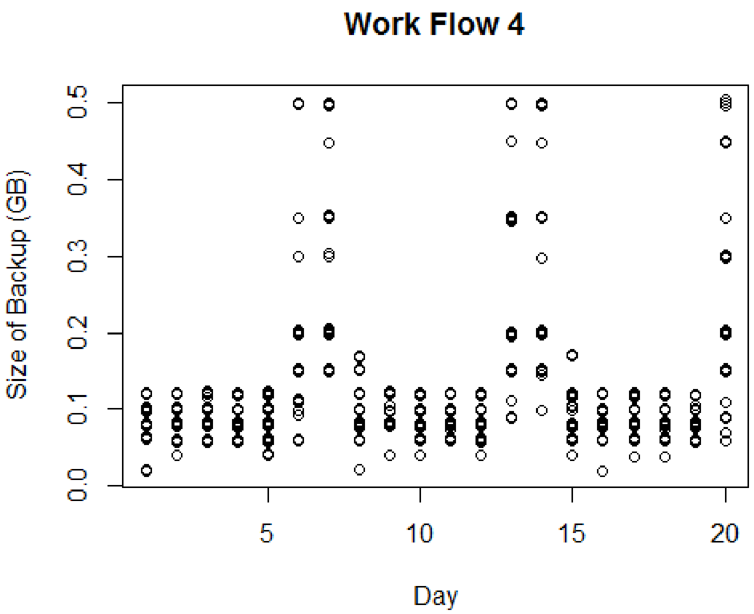
Problem 1






Work Flow 2
Size of Backup (GB)
0.00
0.05
0.10
0.15
0.20
5
10
15
20
Day


Work Flow 3
Size of Backup (GB)
0.00
0.01
0.02
0.03
0.04
5
10
15
20
Day



Each of the 5 work flow types show a pattern that appears to repeat every week. Therefore, there is probably some sort of relationship between something like day of the week and size of the backup.

Problem 2

Part 2a

First, a fitted linear model using the whole data set (result from R):

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.15696 -0.03649 -0.00458  0.02262  0.64911 

Coefficients: (4 not defined because of singularities)
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -8.279e-02  3.730e-03 -22.193   <2e-16 ***
data$Week.                            1.183e-04  1.212e-04   0.976   0.3291    
data$Day.of.WeekMonday                8.200e-02  2.000e-03  41.004   <2e-16 ***
data$Day.of.WeekSaturday              6.838e-02  2.034e-03  33.625   <2e-16 ***
data$Day.of.WeekSunday                6.929e-02  2.041e-03  33.948   <2e-16 ***
data$Day.of.WeekThursday              4.628e-02  2.001e-03  23.133   <2e-16 ***
data$Day.of.WeekTuesday               2.105e-03  1.988e-03   1.059   0.2896    
data$Day.of.WeekWednesday             5.294e-02  2.057e-03  25.736   <2e-16 ***
data$Backup.Start.Time...Hour.of.Day  9.340e-04  7.678e-05  12.165   <2e-16 ***
data$Work.Flow.IDwork_flow_1          3.885e-02  4.189e-03   9.276   <2e-16 ***
data$Work.Flow.IDwork_flow_2          2.379e-03  4.130e-03   0.576   0.5646    
data$Work.Flow.IDwork_flow_3         -7.021e-03  4.135e-03  -1.698   0.0895 .  
data$Work.Flow.IDwork_flow_4          4.039e-02  4.048e-03   9.977   <2e-16 ***
data$File.NameFile_1                  1.244e-03  4.074e-03   0.305   0.7600    
data$File.NameFile_10                -8.777e-04  4.123e-03  -0.213   0.8314    
data$File.NameFile_11                -8.074e-04  4.123e-03  -0.196   0.8447    
data$File.NameFile_12                 1.580e-03  4.040e-03   0.391   0.6957    
data$File.NameFile_13                 2.767e-04  4.046e-03   0.068   0.9455    
data$File.NameFile_14                -8.642e-04  4.046e-03  -0.214   0.8309    
data$File.NameFile_15                -1.434e-03  4.044e-03  -0.355   0.7228    
data$File.NameFile_16                 1.207e-03  4.043e-03   0.299   0.7653    
data$File.NameFile_17                        NA         NA      NA       NA    
data$File.NameFile_18                 1.716e-03  4.023e-03   0.427   0.6697    
data$File.NameFile_19                -5.332e-04  4.023e-03  -0.133   0.8946    
data$File.NameFile_2                  1.548e-03  4.072e-03   0.380   0.7038    
data$File.NameFile_20                 1.391e-04  4.023e-03   0.035   0.9724    
```

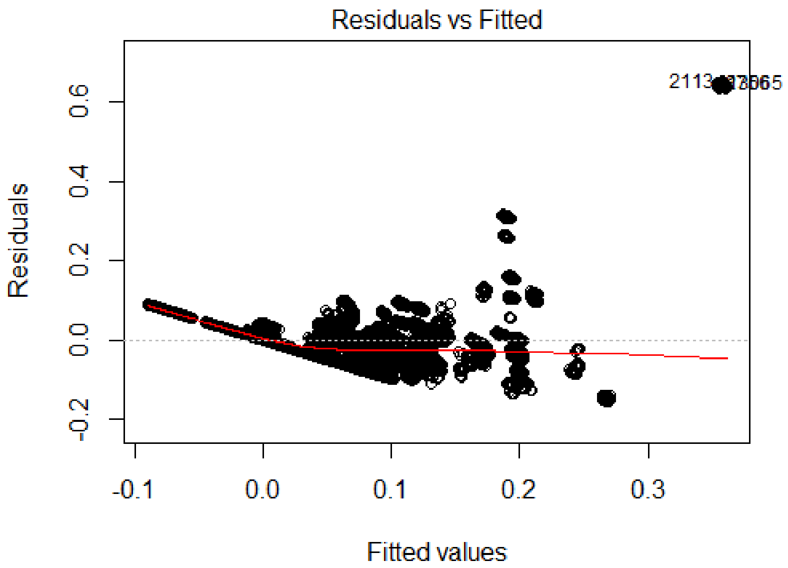
```
data$File.NameFile_21                -1.079e-03  4.023e-03  -0.268   0.7885    
data$File.NameFile_22                -2.477e-03  4.023e-03  -0.616   0.5381    
data$File.NameFile_23                        NA         NA      NA       NA    
data$File.NameFile_24                -1.451e-03  4.023e-03  -0.361   0.7185    
data$File.NameFile_25                -1.010e-03  4.023e-03  -0.251   0.8018    
data$File.NameFile_26                -8.203e-04  4.023e-03  -0.204   0.8384    
data$File.NameFile_27                -5.370e-04  4.023e-03  -0.133   0.8938    
data$File.NameFile_28                -5.043e-04  4.023e-03  -0.125   0.9002    
data$File.NameFile_29                        NA         NA      NA       NA    
data$File.NameFile_3                  1.773e-03  4.072e-03   0.435   0.6633    
data$File.NameFile_4                  5.028e-04  4.072e-03   0.123   0.9017    
data$File.NameFile_5                  1.749e-03  4.075e-03   0.429   0.6678    
data$File.NameFile_6                 -1.245e-03  4.123e-03  -0.302   0.7627    
data$File.NameFile_7                 -9.557e-04  4.123e-03  -0.232   0.8167    
data$File.NameFile_8                 -7.470e-04  4.123e-03  -0.181   0.8562    
data$File.NameFile_9                         NA         NA      NA       NA    
data$Backup.Time..hour.               7.729e-02  7.430e-04 104.028   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07141 on 18549 degrees of freedom
Multiple R-squared:  0.5313,	Adjusted R-squared:  0.5303 
F-statistic: 553.2 on 38 and 18549 DF,  p-value: < 2.2e-16
```

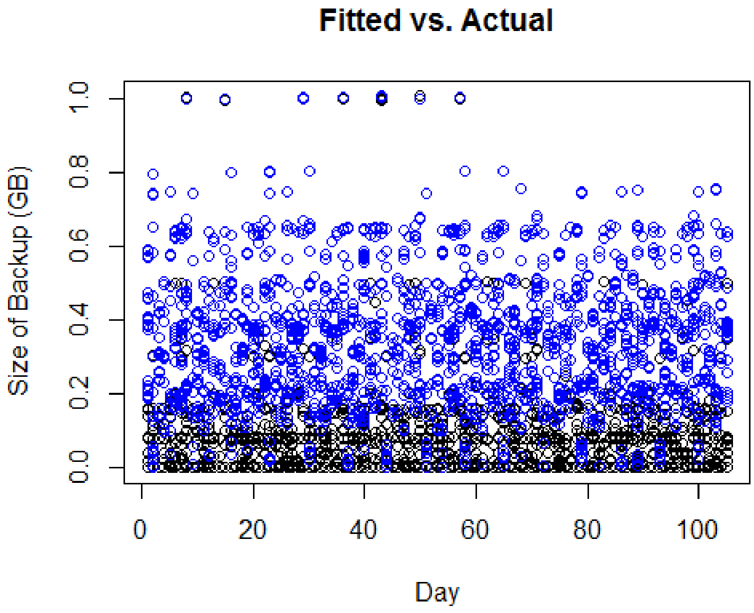
Next, the goal is to perform 10-fold cross validation by taking 10% of the data as a test set and the other 90% as the training set 10 times. The following is the result of one of these cross-validation instances:

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.15625 -0.03631 -0.00466  0.02257  0.65147 

Coefficients: (4 not defined because of singularities)
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           -8.153e-02  3.922e-03 -20.786   <2e-16 ***
train$Week.                            1.917e-04  1.279e-04   1.499   0.1339    
train$Day.of.WeekMonday                8.108e-02  2.107e-03  38.475   <2e-16 ***
train$Day.of.WeekSaturday              6.787e-02  2.135e-03  31.789   <2e-16 ***
train$Day.of.WeekSunday                6.864e-02  2.152e-03  31.899   <2e-16 ***
train$Day.of.WeekThursday              4.585e-02  2.104e-03  21.789   <2e-16 ***
train$Day.of.WeekTuesday               1.680e-03  2.090e-03   0.804   0.4216    
train$Day.of.WeekWednesday             5.241e-02  2.162e-03  24.238   <2e-16 ***
train$Backup.Start.Time...Hour.of.Day  9.152e-04  8.101e-05  11.297   <2e-16 ***
train$Work.Flow.IDwork_flow_1          3.923e-02  4.411e-03   8.894   <2e-16 ***
train$Work.Flow.IDwork_flow_2          1.466e-03  4.326e-03   0.339   0.7347    
train$Work.Flow.IDwork_flow_3         -7.497e-03  4.325e-03  -1.733   0.0831 .  
train$Work.Flow.IDwork_flow_4          4.081e-02  4.263e-03   9.571   <2e-16 ***
train$File.NameFile_1                  1.519e-03  4.295e-03   0.354   0.7237    
train$File.NameFile_10                -3.043e-03  4.329e-03  -0.703   0.4821    
train$File.NameFile_11                -2.543e-03  4.337e-03  -0.586   0.5576    
train$File.NameFile_12                 1.708e-03  4.246e-03   0.402   0.6876    
train$File.NameFile_13                 4.289e-04  4.227e-03   0.101   0.9192    
train$File.NameFile_14                -8.005e-04  4.236e-03  -0.189   0.8501    
train$File.NameFile_15                -1.187e-03  4.238e-03  -0.280   0.7795    
train$File.NameFile_16                 1.266e-03  4.244e-03   0.298   0.7654    
train$File.NameFile_17                        NA         NA      NA       NA    
train$File.NameFile_18                 1.328e-03  4.222e-03   0.315   0.7531    
train$File.NameFile_19                -1.504e-03  4.226e-03  -0.356   0.7219    
train$File.NameFile_2                  4.962e-04  4.315e-03   0.115   0.9085    
train$File.NameFile_20                -3.694e-05  4.203e-03  -0.009   0.9930    
train$File.NameFile_21                -2.247e-03  4.226e-03  -0.532   0.5949    
train$File.NameFile_22                -2.846e-03  4.220e-03  -0.674   0.5001    
train$File.NameFile_23                        NA         NA      NA       NA    
train$File.NameFile_24                -3.196e-03  4.255e-03  -0.751   0.4525    
train$File.NameFile_25                -1.804e-03  4.218e-03  -0.428   0.6689    
train$File.NameFile_26                -1.991e-03  4.247e-03  -0.469   0.6392    
train$File.NameFile_27                -2.247e-03  4.246e-03  -0.529   0.5967    
train$File.NameFile_28                -2.680e-03  4.223e-03  -0.634   0.5258    
train$File.NameFile_29                        NA         NA      NA       NA    
train$File.NameFile_3                  1.291e-03  4.284e-03   0.301   0.7631    
```

```
train$File.NameFile_4                     -2.523e-04  4.293e-03  -0.059   0.9531
train$File.NameFile_5                      1.151e-04  4.282e-03   0.027   0.9786
train$File.NameFile_6                     -1.803e-03  4.349e-03  -0.415   0.6785
train$File.NameFile_7                     -3.181e-03  4.359e-03  -0.730   0.4656
train$File.NameFile_8                     -4.216e-03  4.371e-03  -0.965   0.3348
train$File.NameFile_9                             NA         NA      NA       NA
train$Backup.Time..hour.                   7.694e-02  7.826e-04  98.315   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07139 on 16690 degrees of freedom
Multiple R-squared:  0.5294,	Adjusted R-squared:  0.5283 
F-statistic: 494.1 on 38 and 16690 DF,  p-value: < 2.2e-16
```

For this instance of cross-validation, the RSE was apparently 0.07139. For the test data, the RMSE was 0.071628. Judging from factors such as the R^2 value, perhaps the linear model was not the best model to use, since $R^2 = 0.5294$ is not the greatest. The Fitted vs. Actual plot supports this claim.





For the Fitted vs. Actual plot, the blue points are the fitted values while the black values are the actual values.

The average RMSE of the 10-fold cross validation can also be found for the 10 instances of test/training sets:

0.071628	0.07576351	0.07133117	0.06927778	0.07431938
0.06709536	0.06867551	0.07110193	0.0582345	0.06805916

The average RMSE is found to be about 0.069317589.

From the regression summary output, it can be seen that the effect of the features Day of Week, Backup start time hour of day, work flow ID, and backup time hour all contained coefficients that were significant. This implies that perhaps the features Week and File Name might not be significant enough to be included in the model.

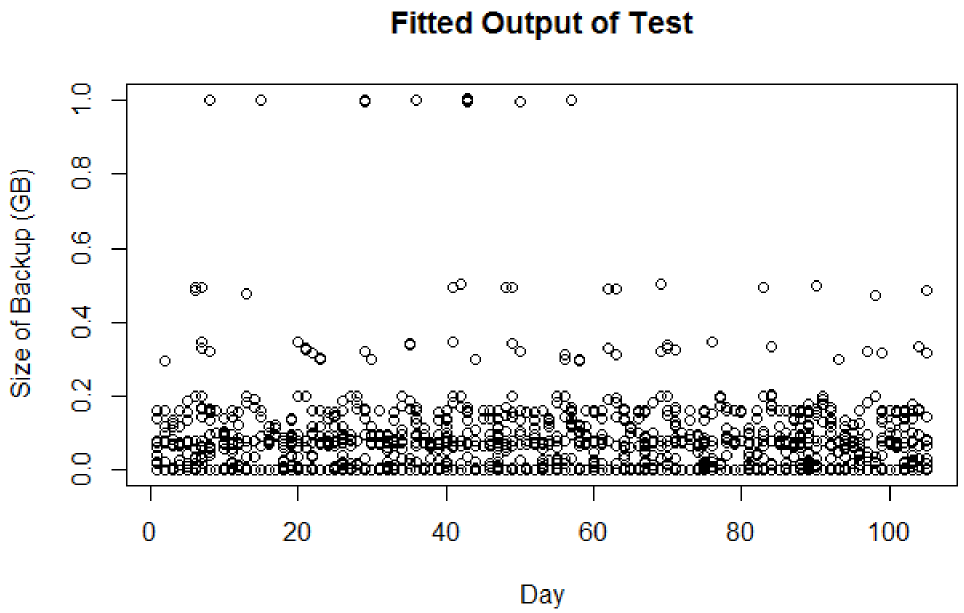
Part 2b

In this part, the randomForest package was installed and used with R. Some RMSE numbers are given below. For training/test set, the same process was used as in part (2a) to get one instance of a partition. The RMSE was found by using `print()` to display the MSE and then taking the square root of that value.

20 trees and all features:

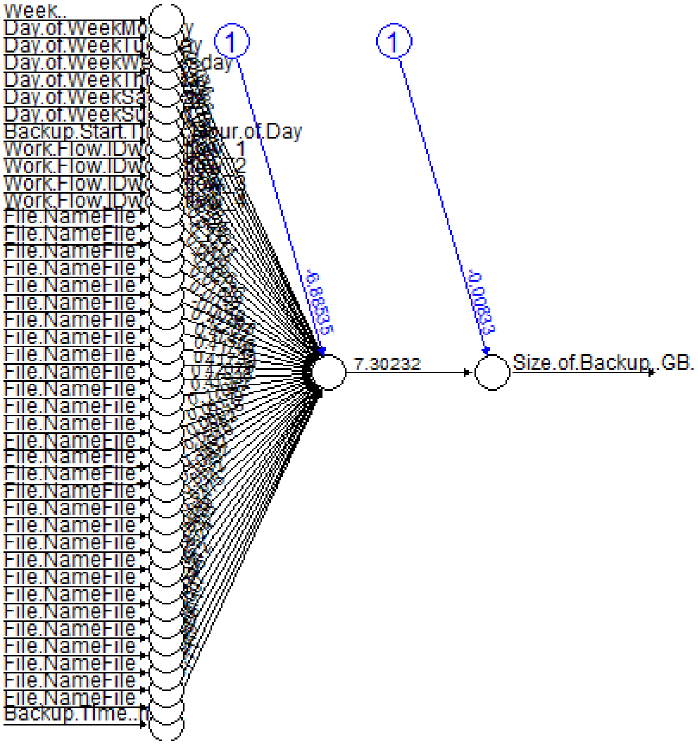
Data	RMSE
Whole Data Set	0.01013244
Training Set	0.01026953
Test Set	0.01050272
50 trees, 6 features	0.009932501
20 trees, 3 features	0.01015536

It appears that the RMSE does not get much better than roughly 0.01. This RMSE is better than that obtained from the linear regression model from before.



There is a similar pattern to part 1 as in every week or so there are a few outliers of large-size backups. The test set is rather small though (10%) so not every week's high values are represented.

Part 2c



The neuralnet package does not accept non-numeric variables, so model.matrix was used to convert the training/test sets. Increasing the amount of neurons in the hidden layer(s) should help improve the performance in regards to the RMSE, but the computation time would increase considerably (the current computation time was already extremely long). Print() returns the following:

```
        Error Reached Threshold Steps
1 13.65010178    0.009958971851 29071
```

Problem 3

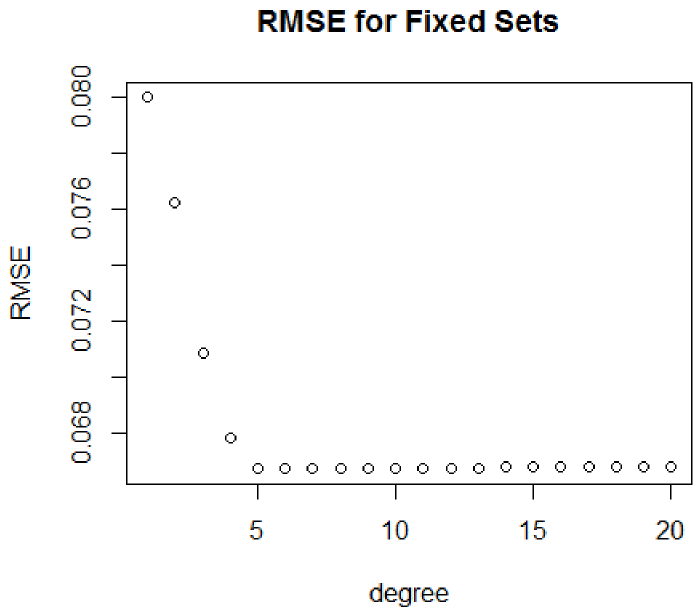
The following chart summarizes the RMSE values for the different workflows:

	Regression Residual Standard Error	R^2	Test RMSE
Workflow 0	0.07187246	0.5276495	0.0718880048
Workflow 1	0.07309647	0.5307583	0.07151021589
Workflow 2	0.07362694	0.5247837	0.07131264461
Workflow 3	0.06878866	0.5359038	0.0723790114
Workflow 4	0.07513672	0.5284148	0.07101819252

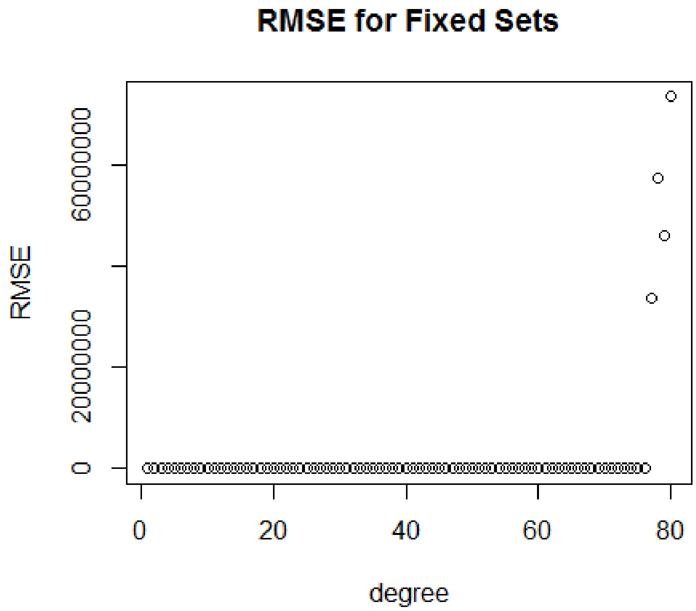
There doesn't seem to be any noticeable improvement when the linear regressions are performed on the workflows separately. This might be because a linear model is not suitable for the data set. The shapes of the patterns observed in problem 1 for each of the work flows were also quite similar.

The next part involved multiple polynomial regression. First, the non-numeric features were converted to numbers using `as.numeric()`.

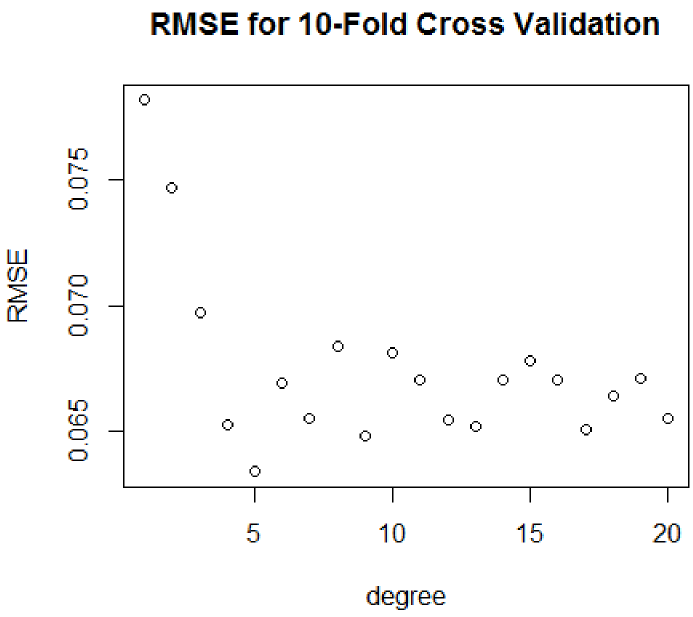
The first figure shows the results from a fixed test/training set.



It appears that after the 5th degree, the RMSE stopped improving and actually got gradually slightly worse.



When the degree axis is expanded, it is apparent that upon approaching the 80th degree, the model's RMSE gets extremely large and therefore a lot worse.



Cross-validation should provide more trials for each degree, potentially making it easier to find the optimal polynomial degree that should be used. Here, it appears that degree 5 is suitable for use.

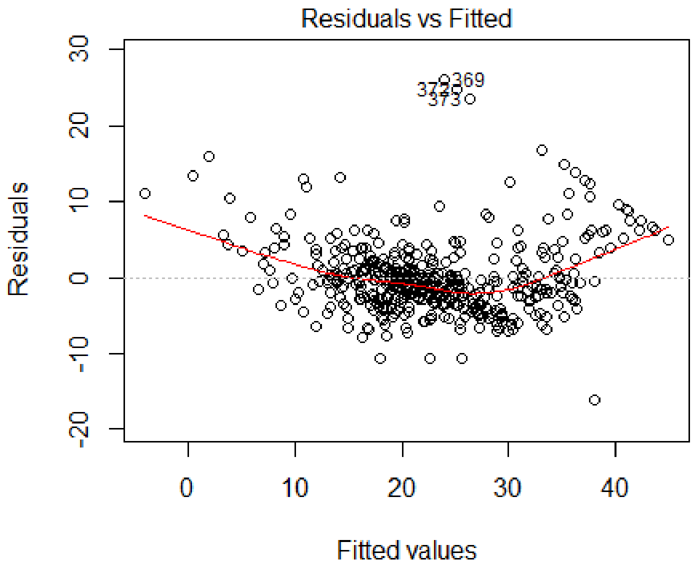
Problem 4

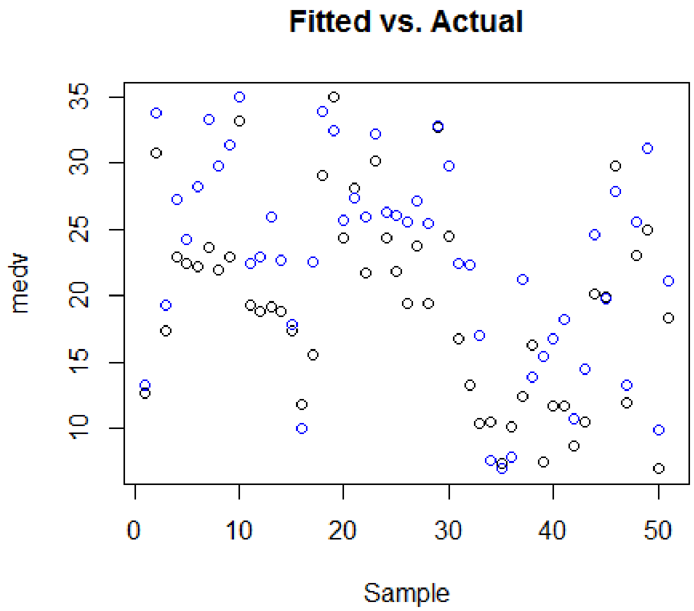
To analyze the significance of the variables, first a linear regression was performed on the whole data set, giving the following output summary:

```
Residuals:
        Min          1Q      Median          3Q         Max
-15.5944739  -2.7297159  -0.5180489   1.7770506  26.1992710

Coefficients:
                 Estimate     Std. Error   t value               Pr(>|t|)
(Intercept)  36.4594883851   5.1034588106   7.14407    0.00000000000328344 ***
crim         -0.1080113578   0.0328649942  -3.28652             0.00108681 **
zn            0.0464204584   0.0137274615   3.38158             0.00077811 ***
indus         0.0205586264   0.0614956890   0.33431             0.73828807
chas          2.6867338193   0.8615797562   3.11838             0.00192503 **
nox         -17.7666112283   3.8197437074  -4.65126    0.00000424564380765 ***
rm            3.8098652068   0.4179252538   9.11614 < 0.000000000000000222 ***
age           0.0006922246   0.0132097820   0.05240             0.95822931
dis          -1.4755668456   0.1994547347  -7.39800    0.00000000000060135 ***
rad           0.3060494790   0.0663464403   4.61290    0.00000507052902269 ***
tax          -0.0123345939   0.0037605364  -3.28001             0.00111164 **
ptratio      -0.9527472317   0.1308267559  -7.28251    0.00000000000130884 ***
b             0.0093116833   0.0026859649   3.46679             0.00057286 ***
lstat        -0.5247583779   0.0507152782 -10.34715 < 0.000000000000000222 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745298 on 492 degrees of freedom
Multiple R-squared:  0.7406427,	Adjusted R-squared:  0.7337897
F-statistic: 108.0767 on 13 and 492 DF,  p-value: < 0.0000000000000022204
```

From the result, the only features that were not significant at any level were “indus” and “age.”

The Residual vs. Fitted and Fitted vs. Actual plots for one iteration of the cross validation:

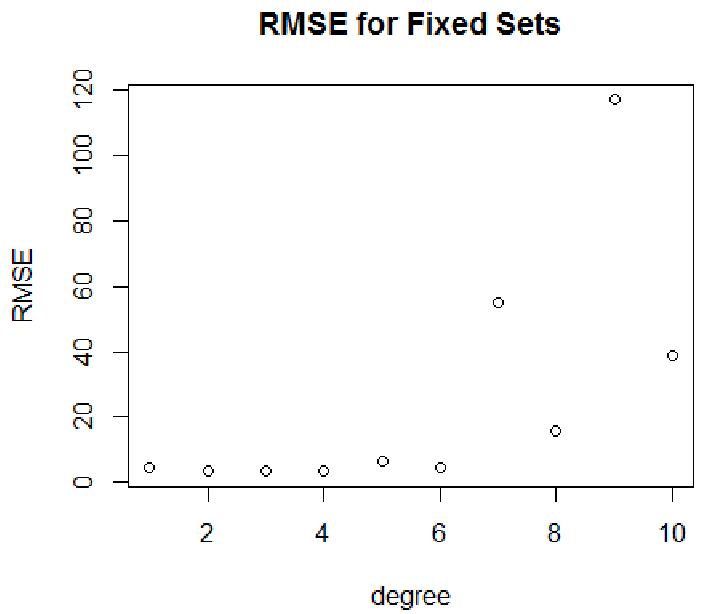




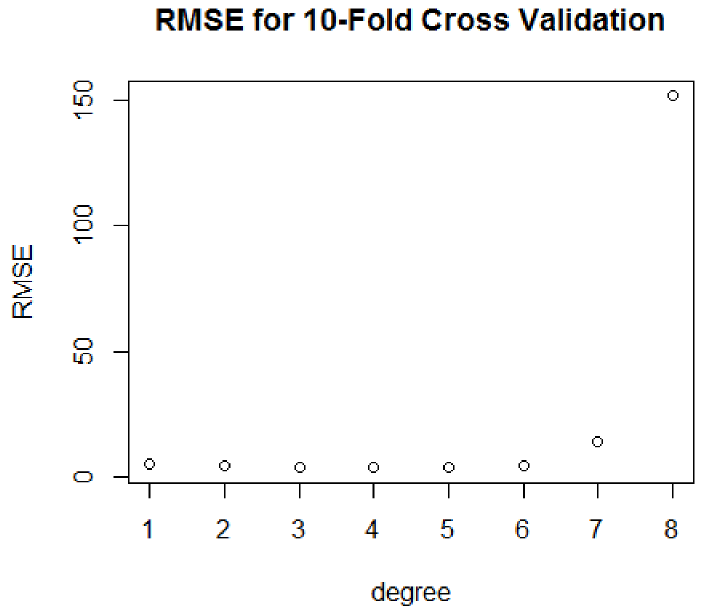
For the Fitted vs. Actual plot, the blue points are the fitted values while the black values are the actual values.

The averaged RMSE from the 10 fold cross validation is found to be 4.400809888.

Next, polynomial regression was performed. First, for a fixed set, the RMSE vs. degree of polynomial fit plot was produced:



Evidently, the model gets a little worse after degree 4, and gets considerably worse after degree 6.



The cross validation results show that the model does indeed start getting a lot worse after about degree 6. The optimal degree appears to be roughly degree 4.

Problem 5

The package “glmnet” was used to perform ridge and lasso regression. The package implements the penalty using this equation:

$$\frac{1-\alpha}{2}||\beta||_2^2+\alpha||\beta||_1$$

Thus, setting the parameter in the fit function “glmnet” alpha = 1 uses Lasso Regularization, while alpha = 0 uses Ridge Regularization. This alpha is not the same as the alpha specified in the instructions; instead, the coefficients of the penalty functions is represented by the lambda parameter, which is set to (0.1,0.01,0.001) as specified.

	RMSE
Lasso	7.632938
Ridge	7.619694