DEC520Q-10B-team27-Team Project Final Report

Ying Guo, Theodore Hector, Zhuoying Lin, Anirudh Reddy, Xiaoshi Zhu

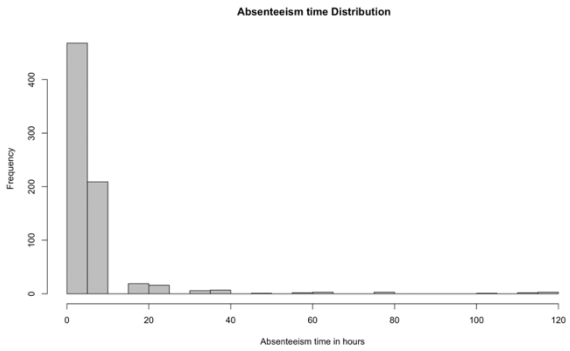
Part 1. Business Understanding

Companies' demand for staff or employees often varies based on business cycles, the overall market, etc. The absence of current employees due to various reasons will add difficulty to the company's deployment of people and the ability to meet the needs of its clientele. Consequently, in our project, we want to investigate what factors could lead to employee absenteeism time. Companies will be able to understand their need for hiring and people deployment better.

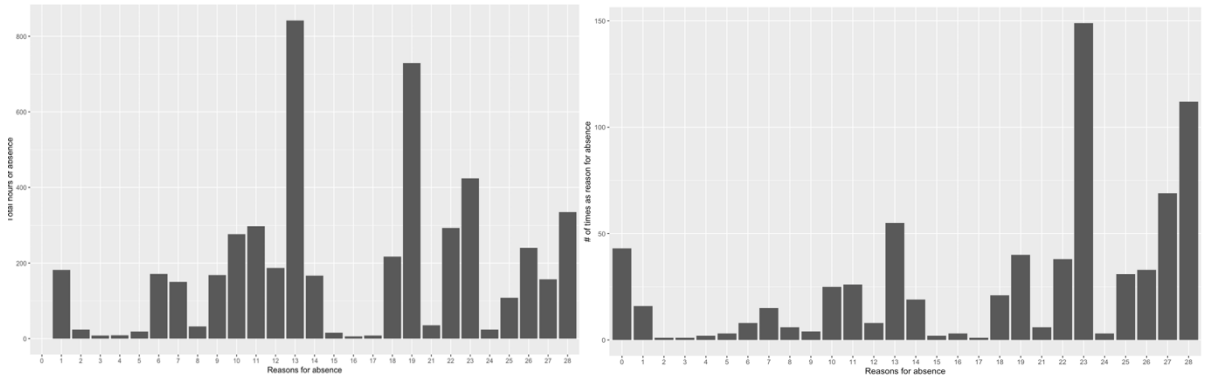
We believe that this would add business value by showing what would cause employees to miss extended hours and potentially implement policies to inoculate these issues. Additionally, the company would be able to investigate the habits of employees and schedule work hours to ensure that workers who have a higher absentee risk are spread out among shifts.

Part 2. Data Understanding

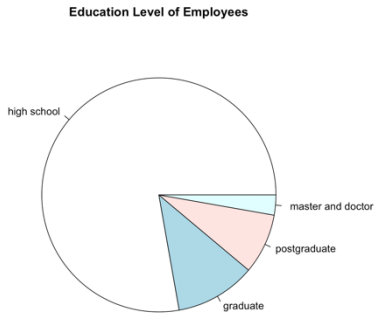
We obtained the dataset we used from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work). It contains records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. The original dataset contains 740 records and 21 variables, including one variable of interest – Absenteeism_time_in_hours, and other employee-specific info such as the number of pets they have, distance to work, ages, etc. We first plot a histogram to see the distribution of Absenteeism time:



We found most people are absent for less than 10 hours each time. Next, we looked at the reasons for absenteeism. The graph on the left shows the total number of hours of absenteeism related to a specific reason, whereas the right shows the total number of times each absenteeism reason is reported. From the first graph, we see that reason 13, having musculoskeletal diseases, relates to the longest total hours in absence. In graph 2, however, we see that the most commonly reported reason is 23, making a blood donation. Since giving blood requires less recovery time than a musculoskeletal disease, we believe that that blood donations account for less of the total absence hours than musculoskeletal diseases.



In terms of education, more than three fourth of employees who reported absenteeism have a high school education. As the level of education increases, fewer people report absenteeism. However, this does not mean that people who are less educated tend to be absent more often, because we do not have data about employees who do not report absenteeism.



Part 3. Data Preparation

We investigated all independent variables in detail and found that the dataset is relatively clean, and there's no missing or wrong value in it. However, we found that some continuous variables should be categorical, such as "ID", "Reason.for.absence", "Month.of.absence", "Day.of.the.week", "Seasons", "Age", "Hit.target", "Disciplinary.failure", "Education", "Social.drinker" and "Social.smoker". Therefore, we converted these values to factor variables.

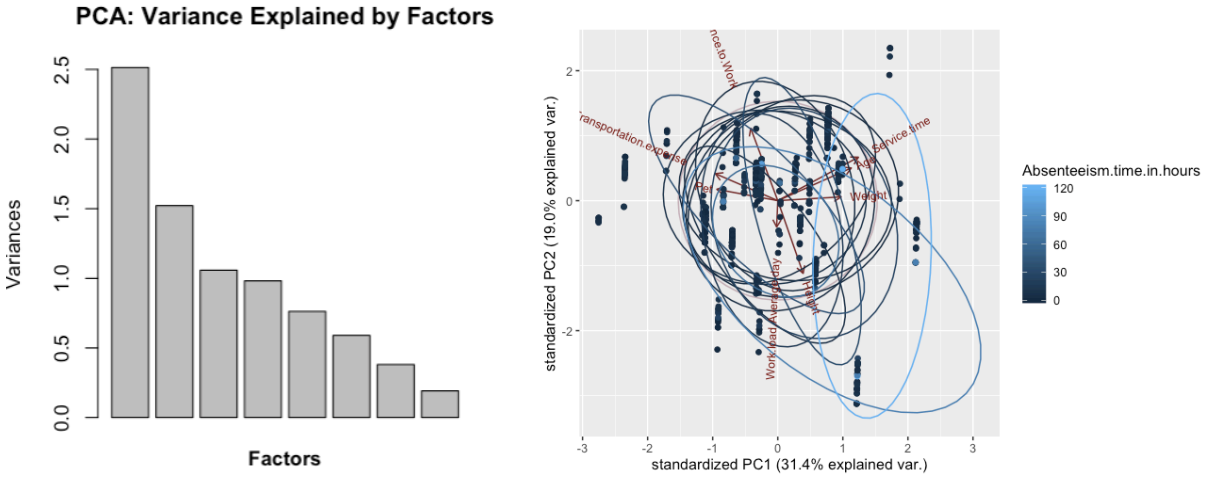
Secondly, we examined the correlation among all pairs of variables and noticed that the "Body.mass.index" is highly correlated with "Weight" (about 0.9). We decided to only include "Weight" in our model. Finally, we looked through the remaining data again to see if there're any outliers but didn't find any obvious one. And since the size of the dataset is not large, we decided not to remove anything and investigate them later after we explored our dataset in more detail.

Part 4. Modeling and Evaluation

PCA

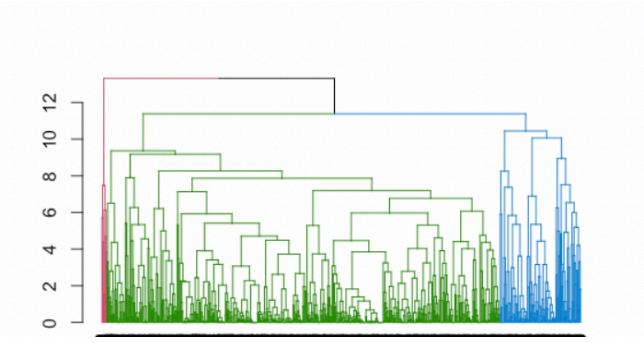
Having obtained a clean dataset, we can analyze the data. Before building models for absenteeism time, we first do unsupervised learning to investigate features of various clusters and to simplify the dataset. We only included continuous variables that work best in PCA. We found that PC1 (Service.time, Age,

Weight, Transportation.expense) explains 31.4% of the total variance, and PC2 (Height and Distance.from.Residence.to.Work) explains 19% of the variance. We formed clusters with respect to absenteeism hours. We found the group with the highest absenteeism hours (lightest blue color) has higher service time, age, and weight. These groups will be informative for factors to include in our predictive model analysis.



Hierarchical Clustering

We perform hierarchical clustering on the dataset using the complete linkage method. We visualize the hierarchical cluster data using a dendrogram plot by coloring the various clusters in the plot. The plot is as follows:



We can observe from the graph that most of the records of the data belong to one large cluster in green; there is a medium-sized cluster in blue, and a small group of records belong to the cluster in red. These employees in cluster 3 are outliers in the dataset.

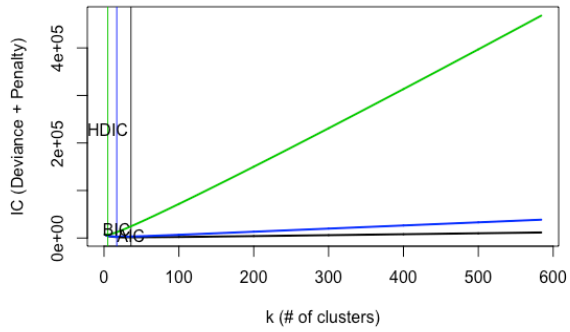
On performing hierarchical clustering on the scaled data and aggregating the data, we get the statistics from the data, as shown in Table 3 of the appendix. We can see that a small group (8 records) of older employees who have been a part of the company for a long time have a very high absentee time on average. Also, the distance to their residence is relatively lower than the other employees. These employees also have no disciplinary failures in the past and have a smaller workload relative to the other employees of the company. These employees have a higher number of children and have an education qualification only up to the graduate level.

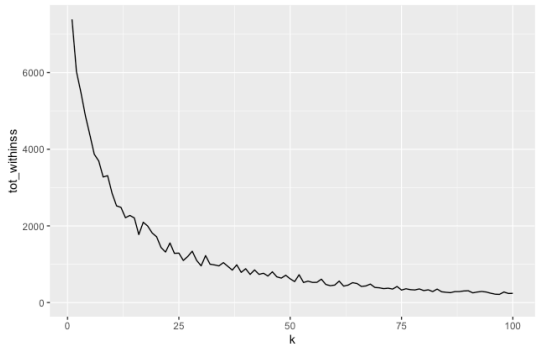
Therefore, with an increase in age, an increase in service time at the company, a decrease in the distance to the residence, a higher number of children, and less workload employee absentee time tends to increase. We can see that education shows a trend where the average absentee time reduces with a higher education level. However, a shift in the hiring standards may have occurred over time and now the companies hire more qualified people. Different hiring standards could imply that education has little or no impact on the absentee hours and that the identifiable trend can be attributed to the standard change.

K-means Clustering

Since K means clustering does not perform well with categorical data, we drop all the variables with categorical data and only consider the continuous variables for the analysis.

Based on the output of the two graphs below, the optimal number of clusters for the dataset to perform k means clustering is 3.





IC vs #of clusters **Elbow Plot**

Selecting the number of centers to be three and performing k means clustering gives us the statistics, as shown in Table 4 of the appendix.

We can see that the increase in age, reduction in transportation cost, a low distance of the company to the residence along with a high service time leads to more absence time in the office. Also, such employees have a higher number of children on average as compared to the remaining population. We can also see that the employees with the most absentee time have the highest workload at the company.

A plot for cluster categorization and a two-dimensional representation of the clusters is in the appendix under the heading Table 5.

Predictive Modeling

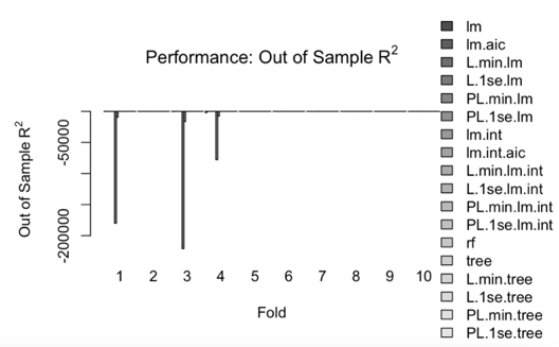
As absenteeism is a continuous variable, we decided linear regression would be an excellent model to use. We first ran a linear regression with all the post-cleaning variables. Unfortunately, the majority of the variables were not significant (see appendix table 1 for detail), and the out of sample R squared were not satisfying. This result was surprising to us as we expected variables such as age or weight would have an impact on absenteeism, as we discovered previously via PCA. To further investigate, we

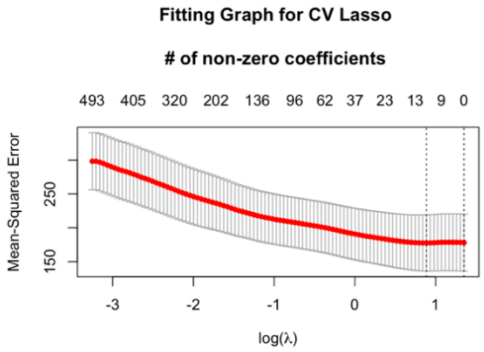
ran stepwise and Lasso for this model. The out of sample R squared for stepwise was no better than that from the original model. The lowest R squared was -186.47.

On the other hand, Lasso and Post Lasso models look much better. However, the models from the first standard error of the mean-squared error had 0 non-zero coefficient, which made it a null model. This coefficient was not helpful to our prediction. Therefore, only Lasso and Post Lasso models from the minimum of the mean-squared error were kept for further comparison.

Although we did not find any variables that have different effects on the relationship between any other variables and absenteeism, we decided to run a linear regression model with interaction. We want to ensure we did not fail to discover any pattern during data exploratory. On the other hand, we wanted to ensure our findings from data exploratory were correct. However, this model is not a suitable model in three ways. First, we had 2207 of variables in total and 678 coefficients, and thus, 1529 variables were utterly redundant. Secondly, none of the variables was significant at the 0.05 level. Lastly, the out of sample R squared of this model was terrible and it was as low as -1.80*10^5, which was much worse than the null model.

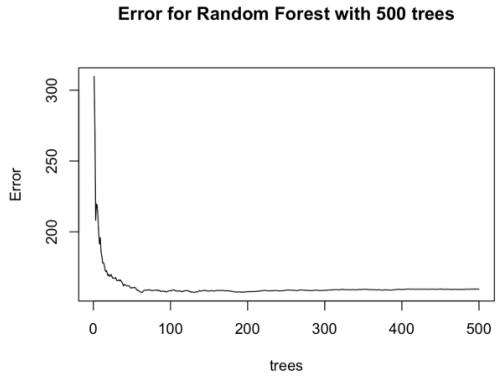
The model was not performing better after we ran stepwise selection and Lasso. Stepwise analysis returns 247 variables and the out of sample R squared is also as low as -1.64*10^4. For Lasso and Post Lasso, the models from the first standard error of the mean-squared error once again selected the null model. Therefore, we did not include these two models in our final comparison. We decided to only keep models from the minimum of the mean-squared error for further investigation.

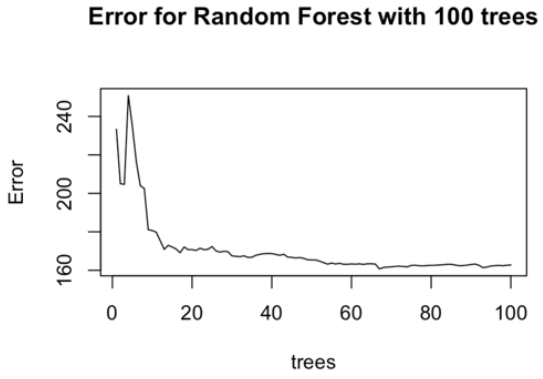




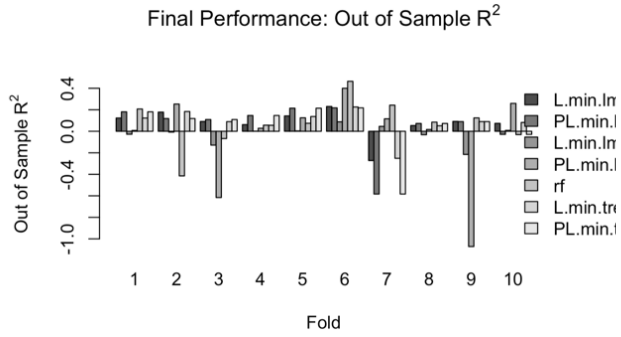
Since the classification tree is easier to understand, we decided to run a classification tree, a classification tree Lasso, and a classification tree Post Lasso. The classification tree was worse than the null model 80% of the times, and once again, models from the first standard error of the mean-squared error selected the null model. Therefore, we only kept models from the minimum of the mean-squared error for further analysis.

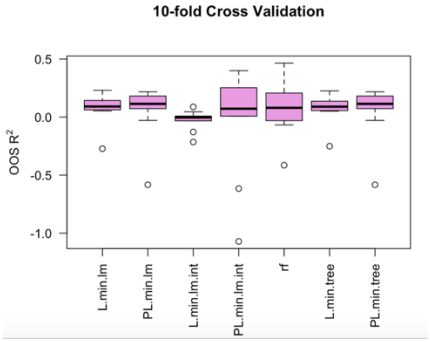
We also ran random forest to average the effect of different trees. Since random forest is less likely to overfit, we did not do Lasso and Post Lasso for it. We started at 500 trees. It was overfitting with 500 trees as error went up after at the 100th tree (see plot below). This discovery makes sense because we only have 740 data. Therefore, we reduce the number of trees to 100. The performance of the random forest was relatively good. Most of the time, the R squared was greater than one, and the highest was 43.57%.



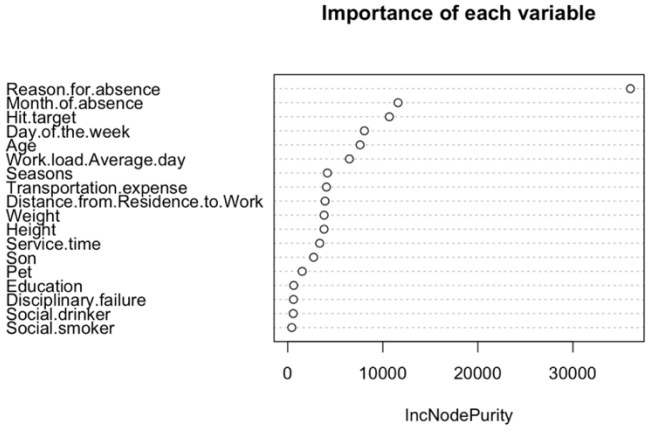


Now we have seven models left for comparison: random forest and Lasso and Post Lasso models for simple linear regression, simple linear regression with interaction, and classification tree. Random forest performed the best for 5 out of 10 times, had the highest mean, and the highest potential out of sample R squared. When we refitted using random forest, the R squared was 0.7316. Therefore, we decided random forest is the most useful model for this dataset.





According to the random forest, “Reason for absence” is the most important feature. “Month of absence”, “Hit Target”, “Day of the week” and “Age” also have relatively significant effects. “Seasons”, “Transportation expense”, “Distance from Residence to Work”, “Weight”, “Height”, and “Service time” have a medium impact on the prediction. The rest of the variables do not have much influence on the prediction of absenteeism time.



Causal Modeling

In addition to predictive modeling, we investigated a causal relationship for absence time. Since Age is statistically significant in all our predictive models, we choose to look at its causal inference on absence time. To isolate the effects of compounders, we used a double selection model. First, we used variables previously selected by Lasso and ran a regression of absence time on those variables except for Age. After getting fitted values, we predicted ages using factors that could affect it, such as service time, education, son, etc. Finally, we ran a regression of absenteeism time on Ages and predicted values from step 1 and step 2. We found Age is significant at 5% level and conclude that a one-year increase in Age would result in 0.19 hour more in absenteeism, holding everything else constant.

Part 5. Deployment

We can draw useful insights for deployment from our analysis. First, we found age has a significant causal effect on absenteeism time, with older people having more absenteeism hours. A company should be prepared for older employees to miss more time than younger ones. Therefore, a company with many older employees can hire more younger people to achieve a balance between work time and experience.

Moreover, we found absence time also varies depending on time. Monday has the highest absenteeism reports and Thursday has the least. This discrepancy might result in unbalanced productivity within a week. As a recommendation, companies can make the process more transparent and allow employees to see when their peers are going to be absent. Employees can then make plans to take days off during times when more people will be working. In this way, the overall productivity is better balanced across days for the company.

Appendix

Table 1: Regression result for simple linear regression model

```
Call:
lm(formula = Absenteeism.time.in.hours ~ ., data = abs_lm)

Residuals:
    Min      1Q  Median      3Q     Max
-33.512  -3.966  -0.565   1.781  99.100

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -65.044621  28.239321  -2.303 0.021562 *
Reason.for.absence1          9.941248   3.707970   2.681 0.007518 **
Reason.for.absence2         25.446989  12.399624   2.052 0.040530 *
Reason.for.absence3          9.281057  12.292194   0.755 0.450489
Reason.for.absence4          4.682115   8.947196   0.523 0.600933
Reason.for.absence5          7.236208   7.294997   0.992 0.321580
Reason.for.absence6         20.027536   4.742487   4.223 2.74e-05 ***
Reason.for.absence7         10.844108   3.726192   2.910 0.003730 **
Reason.for.absence8          8.432693   5.355871   1.574 0.115844
Reason.for.absence9         42.352377   6.437814   6.579 9.49e-11 ***
Reason.for.absence10        11.123423   3.174230   3.504 0.000488 ***
Reason.for.absence11        12.801948   3.161433   4.049 5.73e-05 ***
Reason.for.absence12        24.522361   4.797468   5.112 4.16e-07 ***
Reason.for.absence13        16.723166   2.638327   6.339 4.23e-10 ***
Reason.for.absence14         8.564928   3.411758   2.510 0.012291 *
Reason.for.absence15         4.039664   8.748458   0.462 0.644404
Reason.for.absence16         3.274264   7.451293   0.439 0.660496
Reason.for.absence17         9.680883  12.424099   0.779 0.436133
Reason.for.absence18        10.789176   3.404303   3.169 0.001597 **
Reason.for.absence19        18.433487   2.813590   6.552 1.13e-10 ***
Reason.for.absence21         8.468978   5.422411   1.562 0.118791
Reason.for.absence22         8.839345   2.873820   3.076 0.002184 **
Reason.for.absence23         4.073645   2.274793   1.791 0.073775 .
Reason.for.absence24         8.014129   7.225255   1.109 0.267745
Reason.for.absence25         4.644676   3.074921   1.511 0.131381
Reason.for.absence26         7.040461   2.947150   2.389 0.017171 *
```

```
Reason.for.absence27              6.412735   2.847213   2.252 0.024623 *
Reason.for.absence28              3.338342   2.377806   1.404 0.160789
Month.of.absence1                -4.115026   7.794245  -0.528 0.597702
Month.of.absence2                -3.793252   7.657561  -0.495 0.620506
Month.of.absence3                -2.196639   7.524485  -0.292 0.770428
Month.of.absence4                -2.044713   7.683155  -0.266 0.790220
Month.of.absence5                -4.495881   7.663168  -0.587 0.557609
Month.of.absence6                -0.912145   7.535945  -0.121 0.903696
Month.of.absence7                 3.362425   7.788987   0.432 0.666104
Month.of.absence8                -0.252485   7.905072  -0.032 0.974530
Month.of.absence9                 0.105155   8.006106   0.013 0.989525
Month.of.absence10               -1.887563   8.179876  -0.231 0.817573
Month.of.absence11               -1.265268   8.093456  -0.156 0.875818
Month.of.absence12                0.438700   7.954578   0.055 0.956035
Day.of.the.week3                 -0.503967   1.398239  -0.360 0.718638
Day.of.the.week4                 -1.297647   1.378605  -0.941 0.346899
Day.of.the.week5                 -3.384945   1.466664  -2.308 0.021304 *
Day.of.the.week6                 -2.375938   1.455351  -1.633 0.103027
Seasons2                          3.357135   3.512535   0.956 0.339535
Seasons3                          3.845514   3.135862   1.226 0.220511
Seasons4                          2.818489   3.025000   0.932 0.351807
Transportation.expense            0.012817   0.009422   1.360 0.174185
Distance.from.Residence.to.Work  -0.040562   0.049044  -0.827 0.408501
Service.time                      0.152965   0.192211   0.796 0.426416
Age                               0.175504   0.111245   1.578 0.115116
Work.load.Average.day            -0.015181   0.014586  -1.041 0.298344
Hit.target                        0.081134   0.209389   0.387 0.698522
Disciplinary.failure1                   NA         NA      NA       NA
Education2                       -3.057296   2.252512  -1.357 0.175143
Education3                       -2.848121   1.853798  -1.536 0.124914
Education4                       -6.525724   6.326164  -1.032 0.302653
Son                               0.958245   0.501778   1.910 0.056594 .
Social.drinker1                  -0.350608   1.494402  -0.235 0.814579
Social.smoker1                   -1.274181   2.112384  -0.603 0.546580
Pet                              -0.298717   0.486972  -0.613 0.539806
Weight                           -0.085889   0.053349  -1.610 0.107871
Height                            0.335376   0.107142   3.130 0.001822 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.85 on 678 degrees of freedom
Multiple R-squared:  0.2746,    Adjusted R-squared:  0.2094
F-statistic: 4.208 on 61 and 678 DF,  p-value: < 2.2e-16
```

Table 2: original out of sample R

	lm	lm.aic	L.min.lm	L.1se.lm	PL.min.lm	PL.1se.lm	lm.int	lm.int.aic	L.min.lm.int
1	-1.471861e+02	-1.373663e+02	0.12342995	-0.0006511741	0.18083982	-6.511741e-04	-1.799657e+05	-9.626669e+03	-0.027681394
2	-1.964818e-01	4.769472e-03	0.17819314	-0.0743707194	0.11837111	-7.443278e-02	-6.078656e-01	-9.924245e-02	-0.007714370
3	-5.104020e+01	-1.864696e+02	0.09091658	-0.0017167705	0.10915044	-1.188004e-05	-2.208905e+05	-1.639779e+04	-0.128795975
4	-2.079480e+03	-3.801938e+01	0.06186632	-0.0034970161	0.14670750	-3.212150e-03	-7.781777e+04	-7.645611e+03	0.001144931
5	1.660683e-01	1.626472e-01	0.14275347	-0.0208625902	0.21444770	-2.224980e-02	-3.989465e+00	-5.366714e-02	0.003561641
6	1.033955e-01	2.522195e-01	0.23062466	-0.0175936668	0.21776240	-1.762606e-02	-1.076522e+00	-3.012432e-02	0.087781634
7	-6.972210e-01	-5.351866e-01	-0.27209484	-0.0336022800	-0.58335690	-3.575031e-02	-7.174213e+01	-7.307705e+00	0.045755309
8	3.174044e-02	7.991999e-02	0.05308793	-0.0321645567	0.07239739	-3.415260e-02	-6.220600e-01	-4.317087e-02	-0.031964043
9	-8.882237e-02	2.526884e-02	0.09157381	-0.0108896235	0.08987904	-1.120819e-02	-1.594058e+00	1.929500e-01	-0.215144418
10	-3.503037e-02	5.324773e-03	0.07358129	-0.0150042740	-0.02885868	-1.503536e-02	-3.725081e+01	-1.247999e-01	0.008280284

	L.1se.lm.int	PL.min.lm.int	PL.1se.lm.int	rf	tree	L.min.tree	L.1se.tree	PL.min.tree	PL.1se.tree
1	-0.0006791755	0.007677606	-6.511741e-04	0.20722124	-0.13195499	0.12294157	-0.0006511741	0.18083982	-6.511741e-04
2	-0.0732169456	0.253123821	-7.443278e-02	-0.41477328	-3.97838508	0.18373743	-0.0743707194	0.11837111	-7.443278e-02
3	-0.0092811955	-0.615959960	-1.188004e-05	-0.06762234	-1.34053156	0.08947967	-0.0017167705	0.10915044	-1.188004e-05
4	-0.0032625197	0.028171776	-3.212150e-03	0.05673721	-0.43294513	0.05566537	-0.0034970161	0.14670750	-3.212150e-03
5	-0.0226074530	0.125171712	-2.224980e-02	0.07443003	-0.20581617	0.13710230	-0.0208625902	0.21444770	-2.224980e-02
6	-0.0171922544	0.399945109	-1.762606e-02	0.46484748	0.51717761	0.22569969	-0.0175936668	0.21776240	-1.762606e-02
7	-0.0348808065	0.115847517	-3.575031e-02	0.24274305	-2.28332817	-0.25133089	-0.0336022800	-0.58335690	-3.575031e-02
8	-0.0345005684	0.016732362	-3.415260e-02	0.08502345	-0.22834497	0.05122080	-0.0321645567	0.07239739	-3.415260e-02
9	-0.0258587545	-1.071368918	-1.120819e-02	0.12419508	0.04181276	0.08987796	-0.0108896235	0.08987904	-1.120819e-02
10	-0.0146506034	0.259461711	-1.503536e-02	-0.03062568	-0.21280052	0.08170253	-0.0150042740	-0.02885868	-1.503536e-02

Table 3: Hierarchical Clustering Statistics

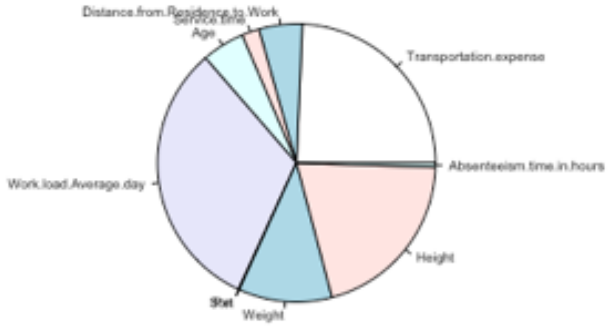
```
> aggregate(absfinal, by=list(abs2$cluster), FUN =mean)
  Group.1  Seasons Transportation.expense Distance.from.Residence.to.Work Service.time      Age
1       1 2.500824              215.9621                        29.53213     12.66557 35.65239
2       2 2.776000              247.7200                        30.84800     11.88800 39.92000
3       3 2.250000              216.2500                        18.12500     14.50000 42.75000
  Work.load.Average.day Disciplinary.failure Education       Son Social.drinker Social.smoker
1              271.6491          0.001647446  1.336079 0.9884679      0.5222405    0.06260297
2              271.2550          0.312000000  1.096000 1.1200000      0.7840000    0.12800000
3              263.1100          0.000000000  1.000000 1.7500000      0.6250000    0.00000000
        Pet   Weight   Height Absenteeism.time.in.hours
1 0.4991763 77.18451 171.9572                 6.194399
2 1.9520000 87.88000 172.6480                 4.128000
3 0.6250000 81.25000 175.7500               106.000000
```

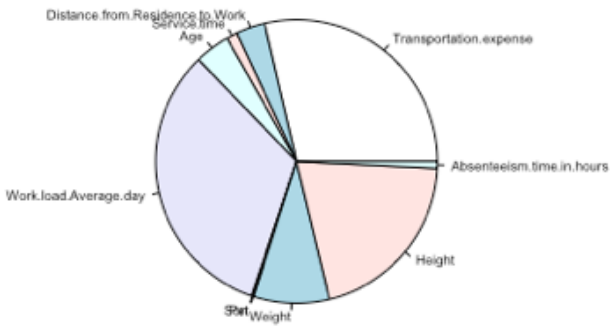
```
> count(abs2, cluster)
# A tibble: 3 x 2
  cluster     n
    <int> <int>
1       1   607
2       2   125
3       3     8
```

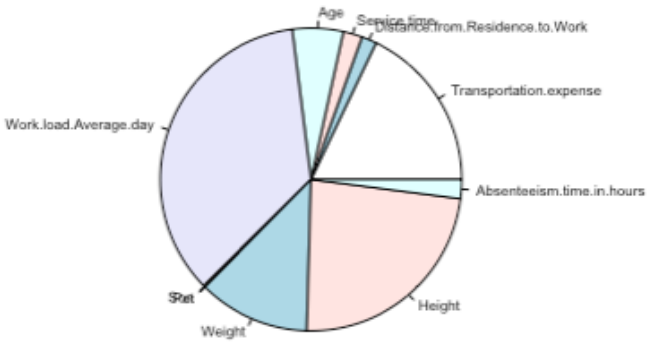
Table 4: K means clustering Statistics

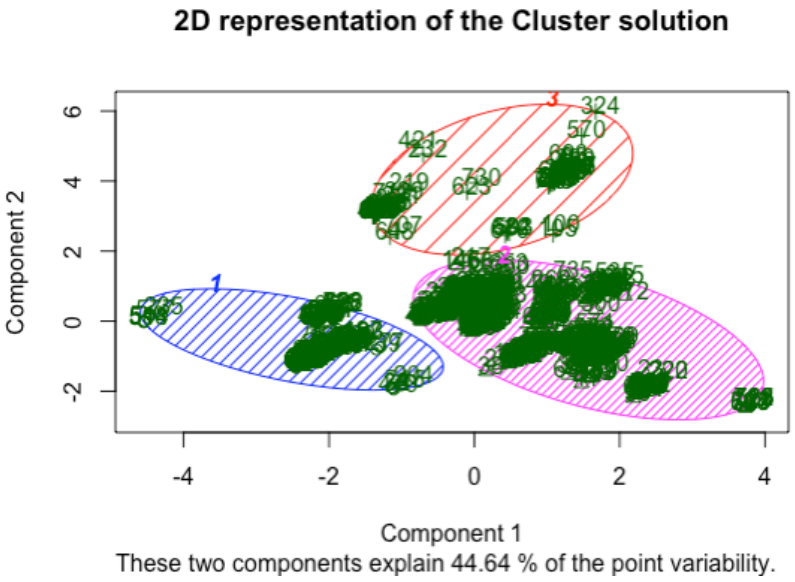
```
> aggregate(abscat, by=list(abs3$clusterkmeans), FUN =mean)
  Group.1 Transportation.expense Distance.from.Residence.to.Work Service.time      Age
1       1              201.7500                        41.79545     16.88636 41.19886
2       2              240.2057                        27.78615     10.46843 33.78819
3       3              141.5753                        12.71233     16.13699 42.90411
  Work.load.Average.day       Son       Pet    Weight   Height Absenteeism.time.in.hours
1              263.3834 0.5113636 0.4261364 90.17045 168.9091                  4.147727
2              273.1659 1.1323829 0.9511202 72.72301 171.3544                  6.553971
3              279.7648 1.4794521 0.1369863 94.64384 184.9589                 16.109589
> count(abs3,clusterkmeans)
# A tibble: 3 x 2
  clusterkmeans     n
          <int> <int>
1             1   176
2             2   491
3             3    73
```

Table 5: Cluster Categorization and 2D representation of the clusters




Distance.from.Residence.to.Work
Service.time
Age
Transportation.expense
Absenteeism.time.in.hours
Work.load.Average.day
Height
Weight


Age
Distance.from.Residence.to.Work
Transportation.expense
Work.load.Average.day
Absenteeism.time.in.hours
Height
Weight



BIBLIOGRAPHY

1. https://www.cognoesis.com/absenteeism-at-work-analysis-prediction/?fbclid=IwAR3VuMdSJM_rb8gNkRFXRM7xASNiBpGUnDoZ6R7sVrNxCj35Pv1QHjn8dOA
2. https://www.datacamp.com
3. https://www.stackoverflow.com
4. UCI Machine Learning Repository

Contributions

Name	Contribution
Teddy Hector	Business Understanding, Deployment and Stepwise Regression
Joy Zhuoying Lin	Predictive Modelling
Xiaoshi Zhu	Data Understanding, Causal Modelling and Deployment
Ying Guo	Principal component Analysis and Data Preparation
Anirudh Reddy	Data Preparation, Cluster Analysis -Hierarchical and K means