PA Model Solution December 2018

This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics.

For this exam, there is a large range of fully satisfactory approaches. This solution illustrates one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.

While the communication and modeling done here is appropriate for the given business problem, candidates should keep in mind that some elements and approaches will not be appropriate for problems given in future sittings. The candidate should always strive to do and write what is relevant to the given business problem.

Failing candidates receive a performance breakdown on the following four topic areas:

Q01 – Executive Summary, Findings and Recommendations

Q02 – Data Exploration and Feature Selection

Q03 – Model Selection and Validation: Decision Tree

Q04 – Model Selection and Validation: Generalized Linear Model

Exam PA December 2018 Project Report Template

Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

To: Peter Stone

From: Exam PA Candidate

Date: December 13, 2018

Executive Summary

Many candidates were too brief in their executive summary. The executive summary should be a snapshot of the entire process, containing the substance of all sections of the full report while using concise, non-technical language. Most candidates listed key factors affecting injury rates, but many did not elaborate on other aspects of the process such as data work.

There are a variety of stylistic practices for executive summaries—elements of style were not important to grading. What was important was being able to trace the through-line from problem to resolution.

An introduction with general commentary on the state of the mining industry is not needed, but it is important to clearly state the business problem to set up the purpose of the report.

The mine workers union has requested two distinct models that predict the rate of injuries per 2000 employee hours for a given mine. They will use this objective, data-driven analysis to independently validate and refine their understanding of the key drivers of mine safety. This will ultimately assist them in developing a simple five-star safety rating that will help their members when choosing where to work and negotiating hazard pay.

Most candidates failed to include a description of the data underlying the analysis, including the data source and type of data. Most candidates failed to discuss data exploration, though some pointed to the body of the report.

The models were built using a dataset from the U.S. Mine Safety and Health Administration (MSHA) from 2013 to 2016, including variables for year, state, mine characteristics, employee count, employee hours, employee activities, and injuries reported. A preliminary analysis led to cleaning the data by removing questionable records and removing the year, primary commodity mined, and state variables. Mine status was revised to address the different terminology between coal and non-coal mines. Initial data exploration showed that the injury rate is higher for underground mines and for coal mines.

The strongest candidates noted where subject matter expertise of the mining union would enhance the analysis.

The mining union may be able to help validate whether the commodity variable sufficiently summarizes the primary commodity mined variable and whether a useful region variable may be created from state data.

Many candidates failed to include a high-level description of the predictive models. The best candidates described the models succinctly, at a high level, in terms that would be understandable to the mine workers union.

Two types of predictive models were used to predict mine injury rates: decision trees and generalized linear models (GLMs). Decision trees are models with a simple, easy-to-interpret structure based on a set of if/then rules that clearly highlight key factors and interactions. GLMs can be used to produce a more comprehensive model, taking all significant variables into account and assessing their relative importance while also producing an easy-to-implement formula to calculate the expected injury rate for a given mine.

Very few candidates identified which model was used for specifying the key factors. The best candidates included clear next steps for modeling that were relevant and meaningful for the mine workers union.

For this analysis, the final GLM provided a more useful and accurate prediction of injury rates, using interaction variables that were identified from the final decision tree. With more time, a decision tree with additional splits may be used to identify other useful interactions that may be added to the GLM to increase predictive power.

Some candidates failed to list the key factors that influence mine safety, the main request of the mining union. Further, of those candidates that listed the key factors that influence mine safety, *many candidates did not clearly state the directional impact of each key factor or convey the relative importance of the different key factors. It is not necessary to list every factor when "key" factors are asked for, but the reader should come away with usable interpretations and meaningful distinctions about the factors.*

Based on the final GLM, the key factors that influence mine safety are:

- The proportion of employee hours spent in the office was a strong indicator of better mine safety, **decreasing** the expected injury rate as the proportion of time spent in office work increases. This is intuitive, as we would expect fewer injuries to happen in an office, and additional office staff may provide effective safety oversight for other workers.
- The proportion of employee hours spent underground was a strong indicator of worse mine safety, **increasing** the expected injury rate as the proportion of time spent underground increases. This is intuitive, as underground mining seems the most hazardous.

Several combinations of type of mine and commodity significantly increased or decreased expected injury rates in comparison to sand and gravel mines, the most common combination in the data:

Type of mine / Commodity	**Increase/decrease on injury rates**
Underground / Coal	+36%
Surface / Stone	+16%
Mill / Stone	+15%
Surface / Metal	+15%
Mill / Metal	-21%
Underground / Stone	-24%
Mill / Nonmetal	-29%
Underground / Nonmetal	-33%

The variation for different types of underground mines is notable and should be discussed further with the mining union.

Finally, the number of employees at the mine was not found to significantly affect injury rates except where time was spent working underground or strip mining. In these mines, having more employees **decreased** injury rates, and the effect was larger as a greater proportion of time was spent on these activities.

Data Exploration and Feature Selection

Many candidates appeared to spend a lot of time on data exploration and feature selection, but only some made their thought process clear and justified their actions. Candidates who did poorly appeared to lack a strategy. While it is not necessary to put into the report every detail of data work done in the code, candidates often showed too little.

As a whole, candidates could have done more to use data visualization to highlight key points of their work in this section. When visualizations are included, however, they should also be interpreted for the reader, something many candidates did not take time to do.

Description and Exploration of Data

> *Most candidates described the data, but few mentioned its applicability or limitations. Those that did mention limitations were often too general, not addressing the specific data and business problem.*

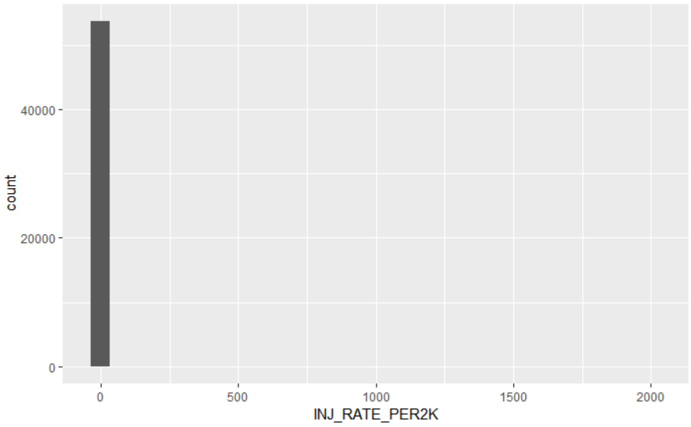
Data from the U.S. Mine Safety and Health Administration (MSHA) ranging from 2013 to 2016 was used for this analysis. The initial data included 53,746 records and 20 variables, including location and nature of mine, employee count and hours, proportions of time worked in various mining operations, and the number of injuries each year. Also, it is not possible to identify trends for individual mines across years, a potentially powerful predictor.

> *For data exploration, candidates often missed looking at the target variable, either the number of injuries or the injury rate. Also, candidates would do well to delineate between univariate and bivariate analysis, being clear why the data is being looked at in a certain way. Both the report and Rmd code were considered in grading all sections of the report.*

A summary of the initial dataset can be found in Table A.1 in the appendix. There are small amounts of missing data in mine status, state, and primary commodity mined, found in only 27 records total. Removing these records should not bias results and so they were deleted.

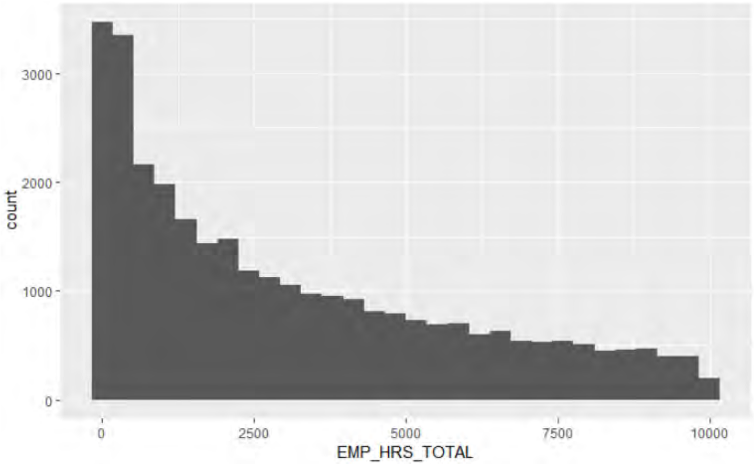
Two variables had a large number of dimensions, which can dilute predictive power. Primary commodity mined had 79 categories, and state had 55 categories. While there may be valuable predictive information contained in these variables, it would require input from the mining union to inform how best to combine these categories into representative groups. For now, the commodity variable appears to capture significant information in the primary commodity mined variable. For state, perhaps the union can suggest suitable groups for analysis. Until then, both variables are dropped from the analysis.

The union asked for a prediction of the injury rate per 2000 employee hours for a given mine. This variable was created and has the following summary:



```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 0.0000    0.0000    0.0000    0.1044    0.0000 2000.0000
```

The maximum of 2000 injuries per employee year is surely due to a reporting error. This mine was found to have 1 employee working 1 hour for the entire year, sustaining an injury in that 1 hour. The 1 employee hour seems unreasonable—the following shows the distribution of employee hours under 10,000 per year:

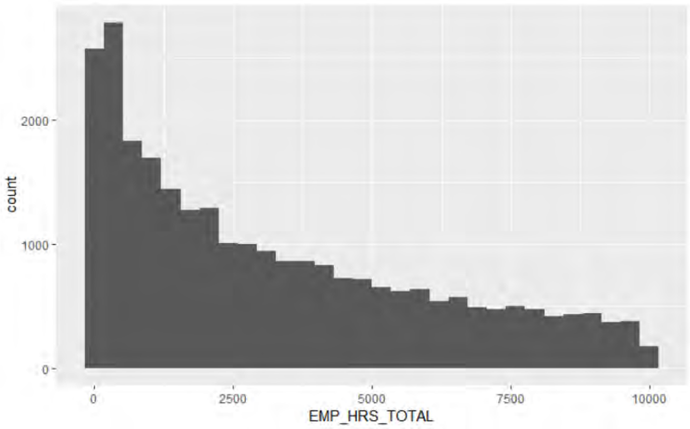


A substantial number of mines have less than 2000 employee hours per mine, the equivalent of 1 full-time employee. Mine status may be affecting this, since some mines are labeled as being closed. The following shows the proportion of mines with less than 2000 employee hours by mine status:

```
        Active        Closed by MSHA   Full-time permanent           Intermittent
    0.07461476            0.17054264            0.02178079             0.48921394
 Non-producing Permanently abandoned     Temporarily closed
    0.34736842            0.55087620            0.43438914
```

Not all candidates explicitly dealt with mine status being different for coal and non-coal mines. There are many approaches to the mine status data, including keeping all the data. The more successful candidates clearly explained the rationale for their actions, relating decisions back to the business problem.

Everything other than “Active” or “Full-time permanent” status have an elevated proportion with low employee hours. While this may be reasonable for “Intermittent”, which in total has 23,039 records, a large portion of the data, there may be reporting inconsistencies for the closed mine types for employee hours. Since the union asked for analysis on “functioning” mines, only “Active,” “Full-time permanent,” and “Intermittent” mines were kept in the data. This removed 5,586 records from the data. The employee hours graph was inspected again:



Some candidates recognized the unusual proportion of very low employee hours. Similar to mine status, there are many justifiable actions that could have been chosen—what is important is not as much the particular action but the justification accompanying it.

While counts of low-hour mines are reduced, there are still a substantial number of mines with unreasonably low employee hours. Because these can skew the injury rate strongly upward, all remaining with fewer than 2000 employee hours (1 full-time employee year) were also removed. This removed an additional 12,013 records. This decision should be discussed with the mining union to confirm that there is not a valid reason for a mine to have such low employee hours.

It was noted that coal and non-coal mines had different designations for mine status. Since "Active" (Coal) and "Full-time permanent" (Non-coal) seem similar in nature, these were combined into a single adjusted status category.

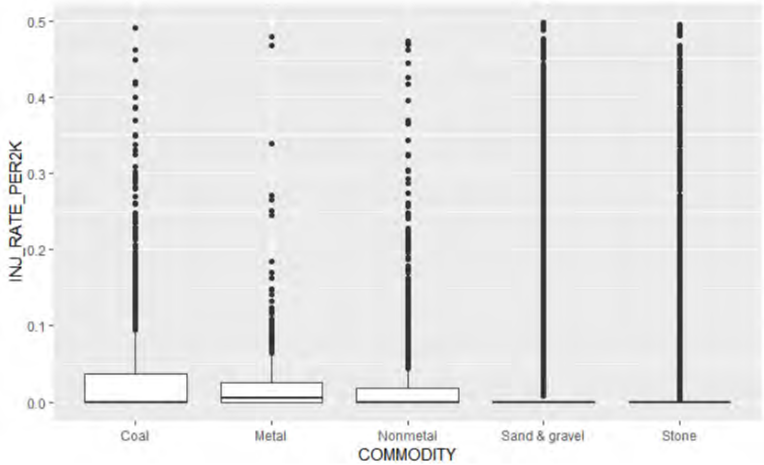
Many candidates removed year, often supported with reasoning, but the strongest candidates checked to make sure a significant trend in injury rates did not need to be considered.

To identify whether year is needed to control for trends in injury rates, the summary for injury rate by year was reviewed:

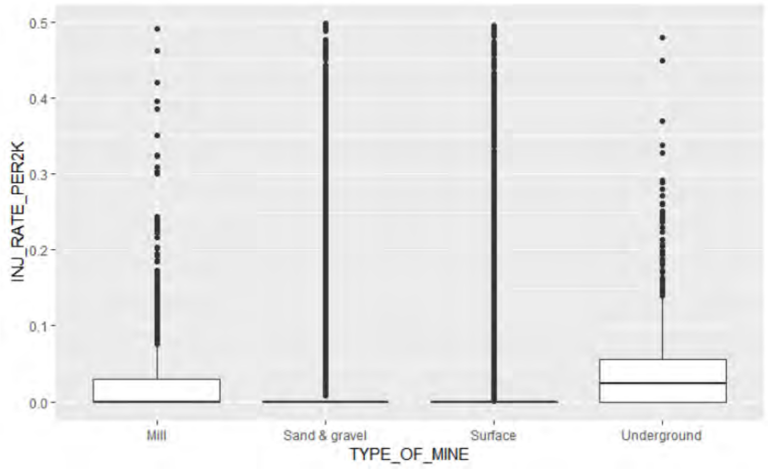
Year	2013	2014	2015	2016
Average injury rate	2.27%	2.38%	2.18%	2.20%

These are in increasing order by year. As can be seen, no significant trend is occurring in average injury rate, so year was dropped as a predictive variable.

To seek likely predictors of injury rate, it was plotted against various predictive variables. Two of the more notable plots are:



This box plot has high injury rates removed to enhance visual differences between categories. By commodity, coal mines have the highest third quartile, indicating higher injury rates overall, while sand & gravel and stone look to have lower injury rates overall.



By type of mine, underground mines appear to be riskiest, as the median in this similarly censored graph is above zero. Sand & gravel and surface types indicate safer working environments.

Few candidates summarized the total effects of data cleaning.

After data cleaning and feature generation, 36,120 records and 16 predictive variables remain, along with the target variable. The most significant data reduction was from removing mines with less than 2000 employee hours per year, a decision that should be reviewed with the mine workers' union. The dropped predictive variables included year, state, and primary commodity mined—the latter two may be used to create useful predictive variables with better understanding of the mining industry.

Model Selection and Validation

Two types of predictive analytic models were applied to predict mine injuries: decision trees and generalized linear models.

Some candidates compared model approaches here, some in the individual model sections, and some in findings. Credit was given for good comparisons regardless of placement. The same is true for splitting train and test data and other elements.

Few candidates explicitly recognized or checked the importance of the use of stratified sampling to improve the effectiveness of the model validation.

For model validation, the data was split into train (75%) and test (25%) sets, using stratified sampling to mitigate the effects that random sampling could have on model outcomes. The target variable means for the train and test sets are consistent. The same partition into train and test sets was used for all models.

Decision Trees

Almost all candidates built a decent-looking decision tree and many tried to interpret it. The interpretations were often literal, with little connection back to injury rates and the business context, though sense checks were usually present. Few candidates used this model to produce a prediction of injury rates, though some reasoned that the tree was more helpful for its interpretation and supporting the GLM than for its own predictions.

Many candidates failed to mention both strengths and weaknesses, along with connection to the business problem for each. High-quality candidates give multiple strengths and weaknesses.

Decision trees will be useful for the analysis for the mining union because they are easy to interpret and explain to non-technical audiences due to the if/then nature of the rules. They also can naturally capture differing characteristics of sub-populations. However, decision trees are prone to overfitting and may not perform well when rating mines in the future. Also, decision trees can change dramatically as new data is added, and having an unstable model would be hard to defend with the mining union and the mines themselves.

Many candidates failed to tie their model validation to an evaluation metric and tie the metric to its business value. It was acceptable to have both a numeric metric and business consideration metric.

Two criteria were used to determine the most appropriate decision tree for this analysis. First, the model needs to be reliable in the future for predicting injury rates. This is measured by maximizing the loglikelihood of the test set. Second, the tree should be of reasonable size for interpretability by miners. This more qualitative criterion is measured by the number of splits and depth of the tree.

The best candidates noted how decision tree pruning was performed. Candidates did not need to use a metric other than the one provided, but if they did, it was expected that it was explained and justified.

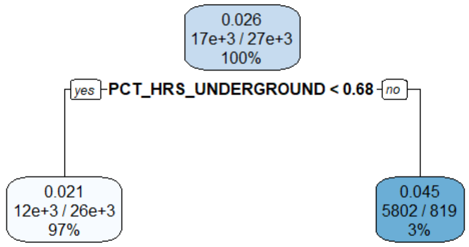
Initially, the decision tree was reduced in size (pruned) by choosing the complexity parameter that minimizes the out-of-sample error validation, or x error. This pruning methodology was later compared to manually choosing the complexity parameter as described below.

Most candidates did not describe the various parameters they tried and how they decided what was better or worse. The strongest candidates had several trees in the write-up and showed how they were working through the model selection and validation process. It is important to include evidence of this process in the code and the report.

A key outcome of this section was to tie the insights back to the business value, as some formulations of models could come up with nonsensical splits. The RMD code has an example of this, where a variable that would not normally be available for prediction is included in the list of predictors. The variable was initially present and then recognized by inspecting the results.

When the decision tree was run on all variables with a minimum complexity parameter of zero, it resulted in a far too complex tree. To address this, an arbitrary minimum complexity parameter of 0.05 was used, along with a minimum bucket size of 25 records. This initial setup (Tree 1) produced the following tree:

Tree 1 tree diagram



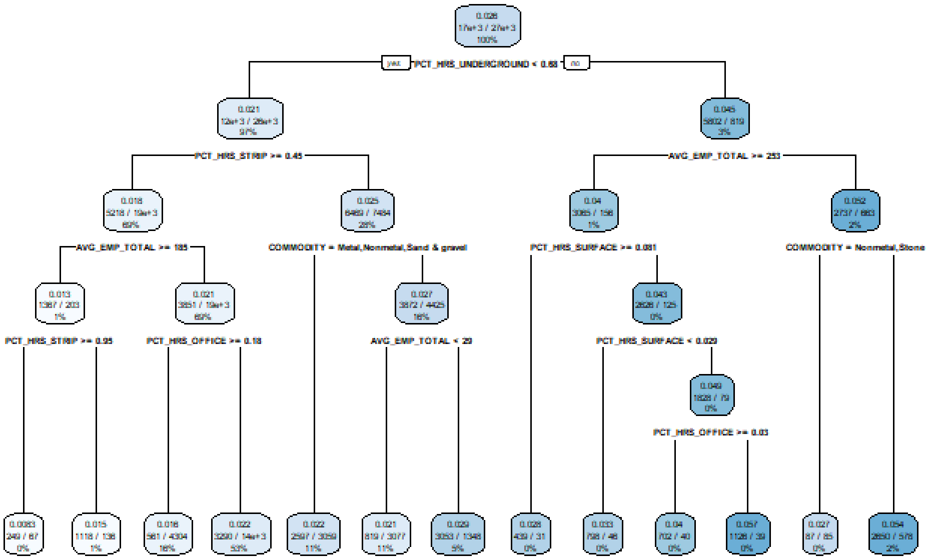
While many candidates ended with some tree that began with this split, only some interpreted it fully and few mentioned how it only segregated 3% of the data.

While this tree only had only one split, it clearly indicates that the percentage of hours underground is the most informative variable. This is intuitive, as working underground is more dangerous than working above ground. However, the union will want to see more detail than this, particularly as it only differentiates injury rates for 3% of mines.

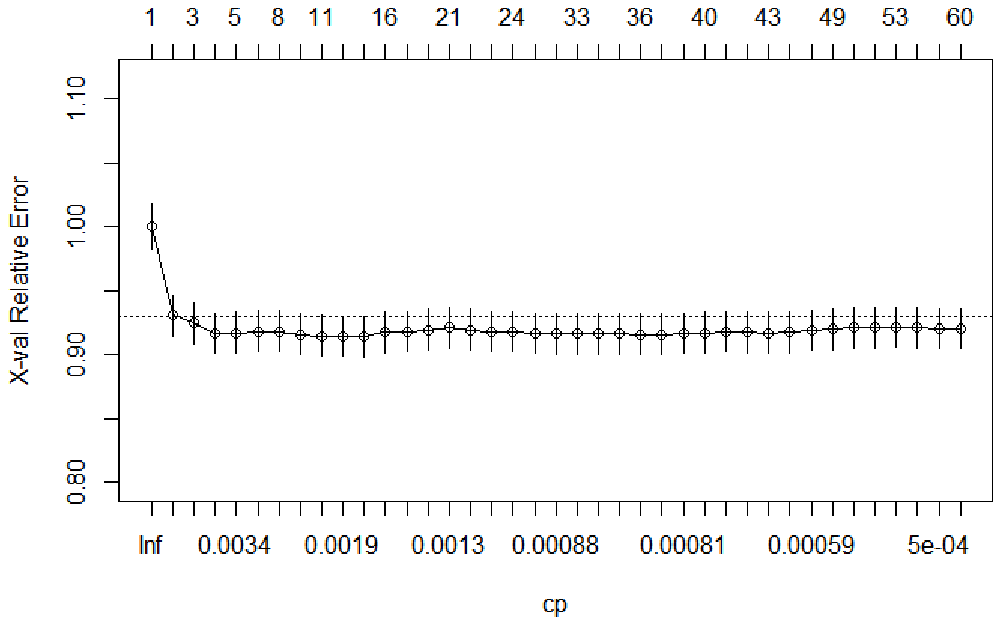
Many candidates adjusted the control parameters, but not many justified it well. Many did say "less complexity" but didn't go deeper than this, causing it to look mostly like trial and error. Few candidates inspected or used the complexity parameter plot included in the Rmd template code to consider how well the cost-complexity pruning was working. Some candidates made effective use of the depth control to achieve the desired size of tree, though this by itself doesn't consider relative accuracy.

The next candidate (Tree 2) was built with a minimum complexity parameter of 0.0005, to allow for larger trees while disallowing the largest possible trees. The pruning methodology is the same, and it resulted in the following tree diagram:

Tree 2 tree diagram

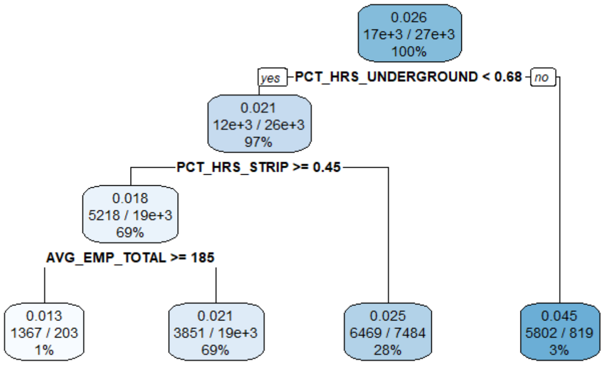


This is too complex of a tree. However, instead of guessing at the parameters, the complexity parameter plot was reviewed:



As the complexity parameter decreases, relative error was initially high, decreased, and then began to increase again, albeit slowly. While the minimum relative error was obtained at 12 splits, a comparable relative error is available with only three splits. As decision trees are prone to overfitting, the tree with 3 splits was selected by manual pruning, using Table A.2 in the appendix as guidance. The relative error between this model and the model with 12 splits is negligible. The following tree (Tree 3) resulted:

Tree 3 tree diagram



This model has three splits, starting with hours underground, then further splitting the largest remaining bucket by hours strip mining and then number of employees. This last split is applied to mines with low hours spent underground and higher hours spent strip mining. It appears that there are interactions between number of employees and hours spent underground or strip mining—this can be explored further in the GLM modeling. Overall, the distinctions seem intuitive and reasonable.

> *To receive full credit, candidates needed to build and comment on at least two models. Loglikelihood on the test data should have been compared between different models. Many candidates instead only compared loglikelihood between the train and test sets on the same model to look for signs of overfitting, missing the primary use of the validation metric, and few among those recognized that scaling was needed to account for differing sample sizes. Some candidates failed to specify what data the loglikelihood applied to.*
>
> *A few candidates made the mistake of refitting the model to the test set. To mimic use of the model for making future predictions, the model fitted to the train set should be used to make predictions from the test set.*
>
> *Other validation metrics such as root mean squared error and mean absolute error could earn credit if the candidate was consistent in its use, but few recognized how the skewed distribution of the target variable compromised these metrics.*
>
> *Few candidates laid out models and validation criteria in an easy-to-digest form.*

The following gives the loglikelihoods for the test data for the three candidate models:

Candidate Model	**Test Set Loglikelihood**
Tree 1	766
Tree 2	935
Tree 3	836

A more positive loglikelihood indicates better predictions on new data. Both Tree 2 and Tree 3 improve upon the too simple Tree 1.

However, while Tree 2 has better predictive power, its high complexity is a detriment. It would be possible to continue refining the balance between predictive power and interpretability by further tweaking the control parameters, but it would be helpful to get feedback from the mining union first. For now, Tree 3 is the best balance of the selection criteria and will be used to inform the GLM modeling.

> *Many candidates made the good recommendation that a random forest would enhance accuracy, either in this section or the recommendations section. Stronger candidates tempered this by also noting how this would come at the cost of interpretability.*

Several of the drawbacks related to decision trees can be overcome by employing a random forest model. Doing so would likely increase the accuracy of predictions but provides little insight other than a ranked list of each variable's predictive power.

Generalized Linear Models (GLMs)

> *For many candidates, it appeared that relatively little time was spent on GLMs. Many candidates followed a rote procedure without describing it well or relating it to the business problem—it was not clear what the point of the process was. Strong candidates explained why they ran the models they ran and what they learned to inform the next model run.*
>
> *While many candidates did well to touch on the use of a Poisson distribution with log link for a count variable, very few discussed simply what a GLM is, nor its methodology or error assumptions. Some candidates did note the possible overdispersion issues ignored by the choice of a Poisson distribution.*

To provide a potentially more predictive model to the mining union with less tendency to overfit, GLMs were also applied. GLMs are a variation of linear regression that allow for non-normal distributions and a functional relationship between the target and a linear function of the predictors. Unlike decision trees, they cannot capture non-linear relationships and are sensitive to the choice of features included.

The GLMs considered use a Poisson distribution with a log link function. For count variables like number of injuries, a Poisson model is appropriate, and a log link function will ensure non-negative predictions. An offset of total employee hours divided by 2000 is used, meaning that the response variable is the number of injuries but the actual value modeled is the injury rate, injuries per 2000 employee hours.

For average number of employees (as well as its interactions), the log of the variable was used as this generally improves model fit for skewed variables. The same cannot be done for seam height due to it having zero as a possible value.

> *Candidates did not handle the fact that sand & gravel was identical in two different variables well. If an error exists, it needs to be addressed. Very few diagnosed the error correctly, partly as many avoided it due to other data cleaning or starting from decision tree. Most candidates that did handle the singularity removed one of the variables entirely instead of trying to retain all the underlying information.*

The initial GLM used all predictive variables in the cleaned data, but this produced multiple errors. One of these was due to sand & gravel being identical in both the type of mine and commodity variables. To remedy this, the two variables were replaced by their interaction, a new variable called mine characteristic which retains all the information.

> *Some candidates recognized the rank-deficient fit due to PCT_HRS_### summing to a constant and the need to remove one of the columns, as transforming to actual hours doesn't help, unless the total hours variable is then dropped. While any could be dropped, justification of which was dropped was often missing. Sometimes this was solved by stepwise selection. Some dropped all of them, losing a lot of valuable predictive information.*

Also, the variables representing proportion of hours spent on different mining activities always sum to 1, causing a different error. Removing one of these removes the problem and essentially adds the removed variable to the baseline. Hours spent strip mining was removed as it is highly associated with the baseline sand and gravel type of mine—other proportional hours will distinguish themselves more clearly in this way, leading to a more predictive model.

> *Few candidates explained how they were selecting superior models. More candidates mentioned the more positive loglikelihood concept given in the template, but many had trouble carrying it through. Some did not take the strong hint on loglikelihood and used RMSE or other metrics, which often earned less credit in this situation.*

To evaluate the predictive power of various GLMs, they are fit using the train set and then the loglikelihood is evaluated on the test set, as was done for the decision trees.

The first valid GLM (GLM 1) includes all the variables as amended to prevent errors without losing information. For the next GLM (GLM 2), stepwise variable selection with AIC was used to remove unimportant variables and reduce the risk of overfitting. Then in a third GLM (GLM 3), potential interactions identified from the final decision tree model were added to the variables used in GLM 2. The final GLM (GLM 4) applies stepwise selection with AIC again to check on the new set of variables.

> *Candidates seldom presented an easy-to-see comparison of model results to assess validation methodology. Instead, comparison of loglikelihoods were cavalier, showing less than mastery of how one validation metric should properly be compared to another. See additional comments in decision trees section. A few candidates attempted to use metrics based on variance explained, a poor choice for this model setup.*

> *Candidates rarely commented on the choice of AIC as the metric or its propensity for retaining more variables than other decision methods.*

These four models produced the following test set loglikelihoods:

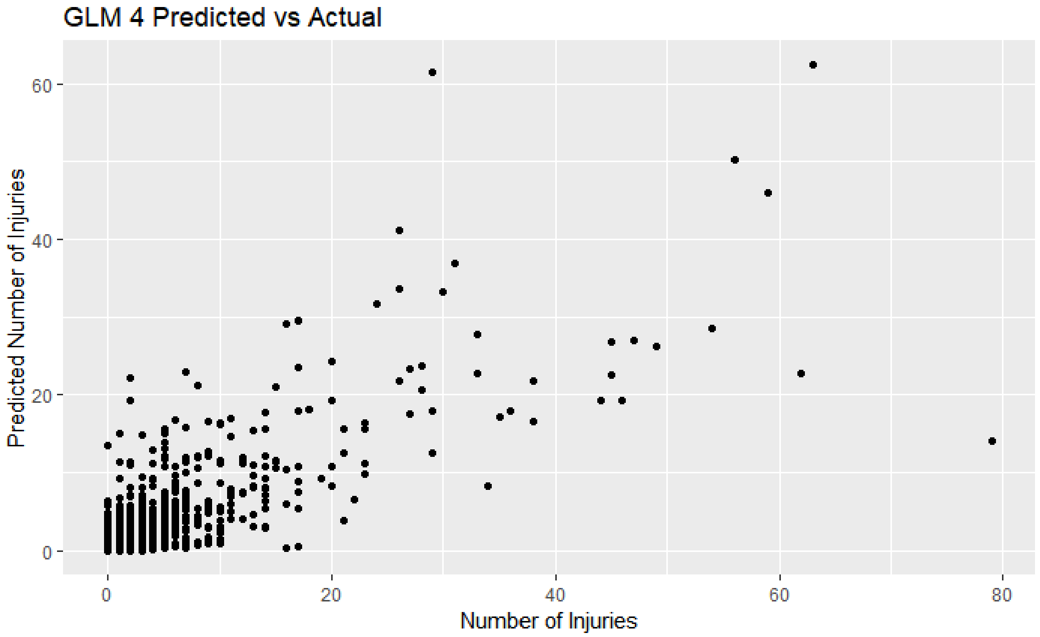
Candidate Model	**Test Set Loglikelihood**
GLM 1	898
GLM 2	900
GLM 3	920
GLM 4	921

The stepwise variable selection removed the adjusted status variable and several of the percent hour variables from GLM 1, but predictive power is only modestly improved. The addition of the two interactions in GLM 3 significantly improves the predictive power, suggesting the interactions are very important. Once these were added, however, hours spent in auger and other surface operations were no longer significant. These are removed in GLM 4, which becomes the recommended model. However, the choice of AIC for the stepwise selection can result in keeping more predictors than some other criteria. Preferences for model complexity should be discussed with the mining union.

> *Candidates had many different good models and selection strategies—this is just one example, though connecting knowledge learned from the decision tree in the GLM is a strong move. Having a clear and well-justified process is key. Some candidates had an overreliance on stepwise regression using AIC without considering its limitations. For example, the drop1 command and the likelihood ratio test could have been used. Given the limited time available, the automated nature of the stepAIC command is an efficient way to proceed.*
>
> *Few candidates performed the valuable step of error analysis, those who did so mostly used q-q plots or residual plots.*

As a final check, the following plot shows the predicted vs. actual number of injuries on the test data using recommended GLM model, GLM 4.

> *Nicely formatted graphs can improve interpretation but are a luxury—more time should be spent on describing process.*



It is expected that the dots of the scatterplot will center around the diagonal with some variance, without leaning one direction or the other. A large proportion of the predicted and actual number of injuries are zero. The resulting figure does not raise any major concerns.

> *Proper interpretation of the coefficients was mostly lacking or poor. Some candidates looked for *** indicated with the p-value and stopped there, noting the factors but not creating a prediction. Stronger candidates related the coefficients back to the predicted injury rate, which is*

more difficult to get from GLM than from the decision tree. Some candidates misinterpreted the GLM as a logistic regression.

Coefficients and p-values for GLM 4 as applied to the full data set are shown in Table A.3 in the appendix.

For a Poisson GLM with a log link function, the baseline expected injury rate is the exponent of the intercept coefficient. For predictive variables, a 1-unit increase of a variable with coefficient beta results in the baseline rate being multiplied by exp(beta). The p-values express the significance of the variables, with smaller being more significant and less than 0.05 being considered statistically significant. A positive effect indicates a positive injury rate and a negative effect indicates a lower injury rate.

Candidates who did interpret their models sometimes did so in the modeling section or in the findings—either is fine.

Discussion of these GLM results is found below in the Findings section.

Findings and Recommendations

Candidates generally did poorly with this section, providing some recommendations but hardly any findings. Many candidates produced models without interpreting them well.

It is not uncommon and perfectly OK for language in the findings and recommendations to appear verbatim in the executive summary if written for both technical and non-technical audiences, as this section also serves as a summary section for portions of the full report. But it is generally not recommended that the two sections be identical—this section should be more robust.

Findings

Some candidates synthesized the outcomes of the decision trees and GLMs, but few wrote well about this, including specifying the ultimate model used. Many candidates went all in on one model without considering what could be gleaned from the other model. Many candidates only discussed decision tree results due to not being able to reconcile errors in the GLMs. The interpretations of these decisions rarely took interactions properly into account, however.

For the test data, the GLM 4 model produced nearly as good a loglikelihood as Tree 2, the most accurate decision tree, and the GLM is more readily interpreted and generalized for the mining union. The decision trees identified the percent hours underground as the most important key factor in determining mine safety, which was confirmed by the GLM. The decision tree also identified an interaction between the percent hours spent underground or strip mining and the total number of employees. These interactions were added to the GLM, which confirmed their significance.

Many candidates used variable names as is when discussing the data, models, and results. While this is acceptable for the technical writing portions of the report, it is distracting and not entirely appropriate for the executive summary.

Some candidates did not run the final model on all available data to provide a more robust prediction for future data. While not mandatory, there should be a good reason given if not done.

Most candidates explained the directional impact of the factors well, though a few failed to realize that positive coefficients for the GLM indicated the negative outcome of more injuries. Many candidates did not provide the common interpretation of multiplicative factors for this type of GLM.

The GLM 4 model was rerun on all data (its coefficients and p-values are in appendix Table A.3) and is used for the following findings.

Based on the final GLM, the key factors that influence mine safety are:

- The proportion of employee hours spent in the office was a strong indicator of better mine safety, **multiplying** the expected injury rate by 0.27 ^ the proportion of time spent in office. For example, spending 5% of the time in the office leads to an injury rate that is $0.27^{0.05}$ = 0.94 times what the rate would be were 0% of the time spent in the office. This is intuitive, as we would expect fewer injuries to happen in an office, and additional office staff may provide effective safety oversight for other workers.
- The proportion of employee hours spent underground was a strong indicator of worse mine safety, **multiplying** the expected injury rate by 2.34 ^ the proportion of time spent underground. This is intuitive, as underground mining seems the most hazardous.

Several combinations of type of mine and commodity significantly increased or decreased expected injury rates in comparison to sand and gravel mines, the most common combination in the data:

Type of mine / Commodity	**Increase/decrease on injury rates**
Underground / Coal	+36%
Surface / Stone	+16%
Mill / Stone	+15%
Surface / Metal	+15%
Mill / Metal	-21%
Underground / Stone	-24%
Mill / Nonmetal	-29%
Underground / Nonmetal	-33%

The variation for different types of underground mines is notable and should be discussed further with the mining union.

The number of employees at the mine was <u>not</u> found to significantly affect injury rates except where time was spent working underground or strip mining. In these mines, having more employees **decreased** injury rates, and the effect was larger as a greater proportion of time was spent on these activities.

Less significant impacts included in the model are higher injury rates for proportion of hours worked for mill prep function and lower injury rates (compared to sand & gravel) for Mill / Coal, Surface / Coal, Surface / Nonmetal, and Underground / Metal mines. Each of these moves affects injury rates by less than 10%. Finally, injury rates are reduced in coal mines by 0.17% for every foot (12 inches) in seam height, and the average seam height among coal mines was 5 feet.

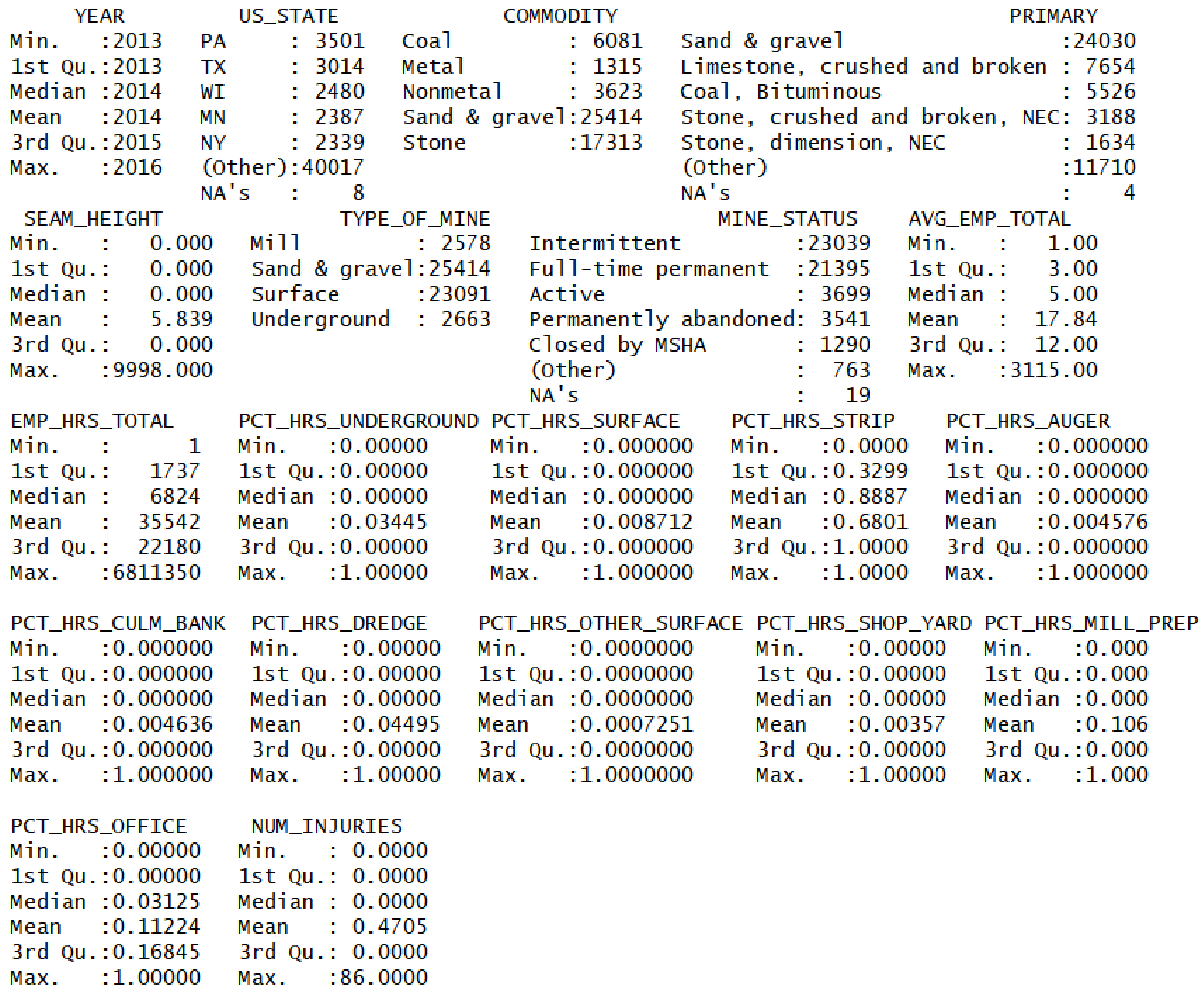
Recommendations

> *Many candidates' recommendations were boilerplate statements without any support or connection to the business problem. Strong candidates avoided a shotgun approach and make a few detailed recommendations specifically considering the business problem and engaging the input of the mining union before moving forward.*
>
> *Many candidates reflexively proposed considering more advanced modeling techniques such as random forests or lasso/ridge regression. While either may indeed provide more predictive power, it is much less clear they will provide better insights for the union. The random forest model may not help identify additional interactions and lasso/ridge may be more difficult to explain.*

The recommended next step is to discuss whether a useful region variable may be created, based on the union's knowledge of whether states in geographic proximity, with similar climates, or with similar levels of mine safety regulation are likely to have similar expected injury rates. They might also consider a cluster analysis to help with that exercise. Also to be discussed is whether the primary commodity mined variable includes important distinctions not found in the commodity variable. After the addition of any new variables based on this discussion, further decision tree modeling with additional splits should be done to identify any other useful interactions that may be added to the GLM to increase predictive power. Finally, this enhanced GLM can be used to create a final formula for expected injury rate by mine, on which the union can quantitatively base their simple five-star rating.

Appendices

Table A.1: Initial data summary



```
      YEAR          US_STATE            COMMODITY                               PRIMARY
 Min.   :2013   PA      : 3501   Coal          : 6081   Sand & gravel                      :24030
 1st Qu.:2013   TX      : 3014   Metal         : 1315   Limestone, crushed and broken : 7654
 Median :2014   WI      : 2480   Nonmetal      : 3623   Coal, Bituminous               : 5526
 Mean   :2014   MN      : 2387   Sand & gravel:25414   Stone, crushed and broken, NEC: 3188
 3rd Qu.:2015   NY      : 2339   Stone         :17313   Stone, dimension, NEC          : 1634
 Max.   :2016   (Other):40017                           (Other)                        :11710
                NA's    :    8                          NA's                           :    4
  SEAM_HEIGHT            TYPE_OF_MINE                   MINE_STATUS       AVG_EMP_TOTAL
 Min.   :   0.000   Mill         : 2578   Intermittent          :23039   Min.   :   1.00
 1st Qu.:   0.000   Sand & gravel:25414   Full-time permanent   :21395   1st Qu.:   3.00
 Median :   0.000   Surface      :23091   Active                : 3699   Median :   5.00
 Mean   :   5.839   Underground  : 2663   Permanently abandoned: 3541   Mean   :  17.84
 3rd Qu.:   0.000                         Closed by MSHA        : 1290   3rd Qu.:  12.00
 Max.   :9998.000                         (Other)               :  763   Max.   :3115.00
                                          NA's                  :   19
 EMP_HRS_TOTAL     PCT_HRS_UNDERGROUND PCT_HRS_SURFACE     PCT_HRS_STRIP    PCT_HRS_AUGER
 Min.   :      1   Min.   :0.00000     Min.   :0.000000   Min.   :0.0000   Min.   :0.000000
 1st Qu.:   1737   1st Qu.:0.00000     1st Qu.:0.000000   1st Qu.:0.3299   1st Qu.:0.000000
 Median :   6824   Median :0.00000     Median :0.000000   Median :0.8887   Median :0.000000
 Mean   :  35542   Mean   :0.03445     Mean   :0.008712   Mean   :0.6801   Mean   :0.004576
 3rd Qu.:  22180   3rd Qu.:0.00000     3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.000000
 Max.   :6811350   Max.   :1.00000     Max.   :1.000000   Max.   :1.0000   Max.   :1.000000

 PCT_HRS_CULM_BANK  PCT_HRS_DREDGE     PCT_HRS_OTHER_SURFACE PCT_HRS_SHOP_YARD PCT_HRS_MILL_PREP
 Min.   :0.000000   Min.   :0.00000   Min.   :0.0000000     Min.   :0.00000   Min.   :0.000
 1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.0000000     1st Qu.:0.00000   1st Qu.:0.000
 Median :0.000000   Median :0.00000   Median :0.0000000     Median :0.00000   Median :0.000
 Mean   :0.004636   Mean   :0.04495   Mean   :0.0007251     Mean   :0.00357   Mean   :0.106
 3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.0000000     3rd Qu.:0.00000   3rd Qu.:0.000
 Max.   :1.000000   Max.   :1.00000   Max.   :1.0000000     Max.   :1.00000   Max.   :1.000

 PCT_HRS_OFFICE     NUM_INJURIES
 Min.   :0.00000   Min.   : 0.0000
 1st Qu.:0.00000   1st Qu.: 0.0000
 Median :0.03125   Median : 0.0000
 Mean   :0.11224   Mean   : 0.4705
 3rd Qu.:0.16845   3rd Qu.: 0.0000
 Max.   :1.00000   Max.   :86.0000
```

Table A.2: Tree 2 complexity parameter table

```
           CP nsplit rel error  xerror     xstd
1  0.07393056      0   1.00000 1.00047 0.017856
2  0.01062643      1   0.92607 0.93067 0.015986
3  0.00848722      2   0.91544 0.92475 0.016024
4  0.00386348      3   0.90696 0.91681 0.015525 <- Manual pruning for tree 3
5  0.00299387      4   0.90309 0.91717 0.015586
6  0.00298042      5   0.90010 0.91828 0.015618
7  0.00220894      7   0.89414 0.91825 0.015772
8  0.00208479      9   0.88972 0.91566 0.015770
9  0.00202295     10   0.88763 0.91458 0.015706
10 0.00187531     11   0.88561 0.91448 0.015729
11 0.00186303     12   0.88374 0.91389 0.015740 <- Automatic pruning, tree 2
12 0.00148571     15   0.87815 0.91772 0.015861
13 0.00139181     16   0.87666 0.91784 0.015736
14 0.00134616     19   0.87249 0.91923 0.015789
15 0.00127715     20   0.87114 0.92094 0.015830
```

Table A.3: Final GLM Model Coefficients

Variable	Estimate	P-Value
(Intercept)	-3.522	0.000
SEAM_HEIGHT (per 12 inches)	-0.0017	0.012
PCT_HRS_UNDERGROUND	0.851	0.000
PCT_HRS_MILL_PREP	0.061	0.283
PCT_HRS_OFFICE	-1.323	0.000
MINE_CHARMill Coal	-0.064	0.244
MINE_CHARMill Metal	-0.242	0.000
MINE_CHARMill Nonmetal	-0.345	0.000
MINE_CHARMill Stone	0.142	0.013
MINE_CHARSurface Coal	-0.085	0.062
MINE_CHARSurface Metal	0.142	0.001
MINE_CHARSurface Nonmetal	-0.026	0.528
MINE_CHARSurface Stone	0.150	0.000
MINE_CHARUnderground Coal	0.310	0.000
MINE_CHARUnderground Metal	-0.057	0.480
MINE_CHARUnderground Nonmetal	-0.398	0.000
MINE_CHARUnderground Stone	-0.278	0.000
LOG_AVG_EMP_TOTAL	0.007	0.527
LOG_AVG_EMP_TOTAL:PCT_HRS_UNDERGROUND	-0.123	0.000
LOG_AVG_EMP_TOTAL:PCT_HRS_STRIP	-0.139	0.000

Note: The coefficient for SEAM_HEIGHT is the estimate from the analysis multiplied by 12 to convert inches to feet.