Predicting the Environmental Damage of Pipeline Accidents

Michael Church Carson

December 16, 2021

TABLE OF CONTENTS

1 INTRODUCTION

In practice, oil pipelines connect one place to another, but these days they seem better at driving us apart. The use of pipelines to transport oil all over North America is becoming more and more controversial. On one hand, they are a relatively safe and efficient way to transport fossil fuels, a resource that we are still hopelessly reliant on. On the other hand, they are expensive, enable our fossil fuel addiction, and when they fail, the consequences for the environment can be catastrophic.[1,2] If a pipeline leaks, breaks, or explodes, the damage to the environment is not only severe, but also unpredictable. Pipeline accidents happen quickly and without warning. As a result, it is difficult to determine when or where an accident will happen, and what the significance of the environmental damage will be.

The cause of a pipeline accident is extremely difficult to predict because there are so many factors (weather, temperature, falling rock, etc.) that can lead to an accident, and one could argue that even if the accidents could be predicted, preventing them would be impossible. Some things are beyond human control—accidents happen. Given the unpredictability and inevitability of pipeline accidents, rather than predict if an accident happens, this report will provide a method to predict the severity of a pipeline accident soon after it happens. More specifically, this report will predict the severity of the environmental damage that a pipeline accident will cause.

Using the U.S. Department of Transport's Pipeline and Hazardous Materials Safety Administration's (PHMSA) pipeline accident reports from 2010 to 2017, this report will provide a descriptive and predictive analysis of pipeline accidents in the United States, and ultimately deliver a statistical model for predicting the severity of the environmental damage a pipeline accident will cause. This report features a description of the data in the pipeline accident reports, an explanation of the methods used to select and design the statistical model, and a discussion of the results of this work and its implications.

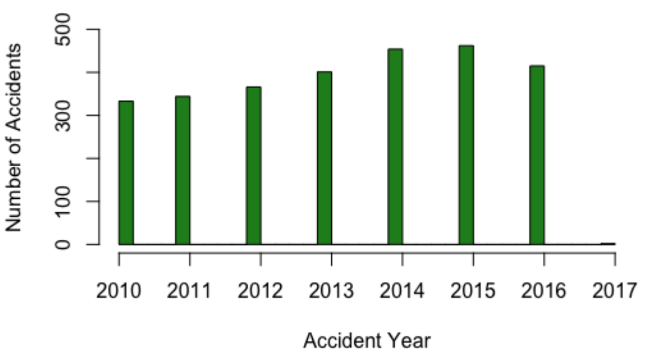
[1] Douglas Bessette, Michelle Rutty, Grant Gunn, Volodymyr Tarabara, Robert Richardson, The perceived risk of the Line 5 Pipeline and spills under ice, Journal of Great Lakes Research, Volume 47, Issue 1, 2021, Pages 226-235, ISSN 0380-1330, https://doi.org/10.1016/j.jglr.2020.12.002.

[2] Auburn University College of Liberal Arts, & Phillips, B. (2017). *Oil Pipelines and Spills - Climate, Energy, and Society - College of Liberal Arts - Auburn University*. Auburn University College of Liberal Arts. https://cla.auburn.edu/ces/energy/oil-pipelines-and-spills/

2 Data Description

2.1 An Overview of The Pipeline Accidents

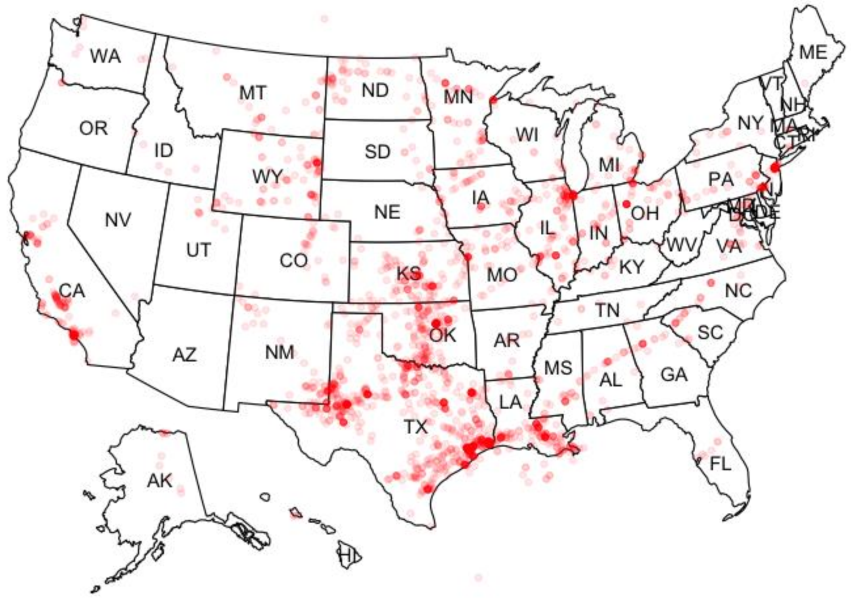
Figure 1: Accidents Per Year (2010-2017)



Between 2010 and 2017, the PHMSA recorded 2,795 pipeline accidents in the United States (on or offshore). As Figure 1 shows, the number of accidents increased year after year from 2010 to 2015, and only dropped slightly in 2016. The number of accidents in 2017 is unclear because the dataset only contains accidents up until January 2017. Nevertheless, pipeline accidents remained at least an issue (if not a growing issue) over this period. The locations of pipeline accidents during this period were distributed across the United States (red dots on Figure 2). Figure 2 shows how the accidents are concentrated in states like Texas, Oklahoma, and Kansas (Appendix 6.2 displays the top 15 states for accidents), but accidents are still a threat to most of the country. This initial exploration of the accidents confirms that the threat they pose to the environment is consistent and widespread. Accordingly, the severity of the environmental damage they can cause is worth investigating and predicting.

Figure 2: US Pipeline Accidents

Source: Dept. of Transport, 2010-2017



2.2 Selecting the dependent and independent variables

The first challenge in predicting the severity of the environmental damage from a pipeline accident is deciding how to quantify the damage. Environmental damage is objectively difficult to measure, and the pipeline accident reports do not include details on how the environment was affected by each accident; however, they do include the total environmental remediation costs associated with each accident. Environmental remediation costs are the costs associated with repairing (to the extent possible) the environmental damage caused by an accident. Consequently, environmental remediation costs present a useful proxy for the damage that an accident does to the environment. For the purposes of this report, environmental remediation costs will be the outcome that the model will predict as a way of *indirectly* predicting the damage that a pipeline accident will cause to the environment.

After pre-processing the data from the accident reports to remove missing values and labelling variables, there are fourteen variables remaining in the dataset that can be used as independent (or predictor) variables for environmental damage.[3] These variables are all characteristics of the pipeline (e.g., operator company, substance transported, above or below ground, etc.), or characteristics of the accident itself (e.g., was there an explosion, did the accident shutdown the pipeline, etc.). Please see Appendix 6.1 for a complete data dictionary with descriptions of the dependent (outcome) variable and the fourteen potential independent (predictor) variables.

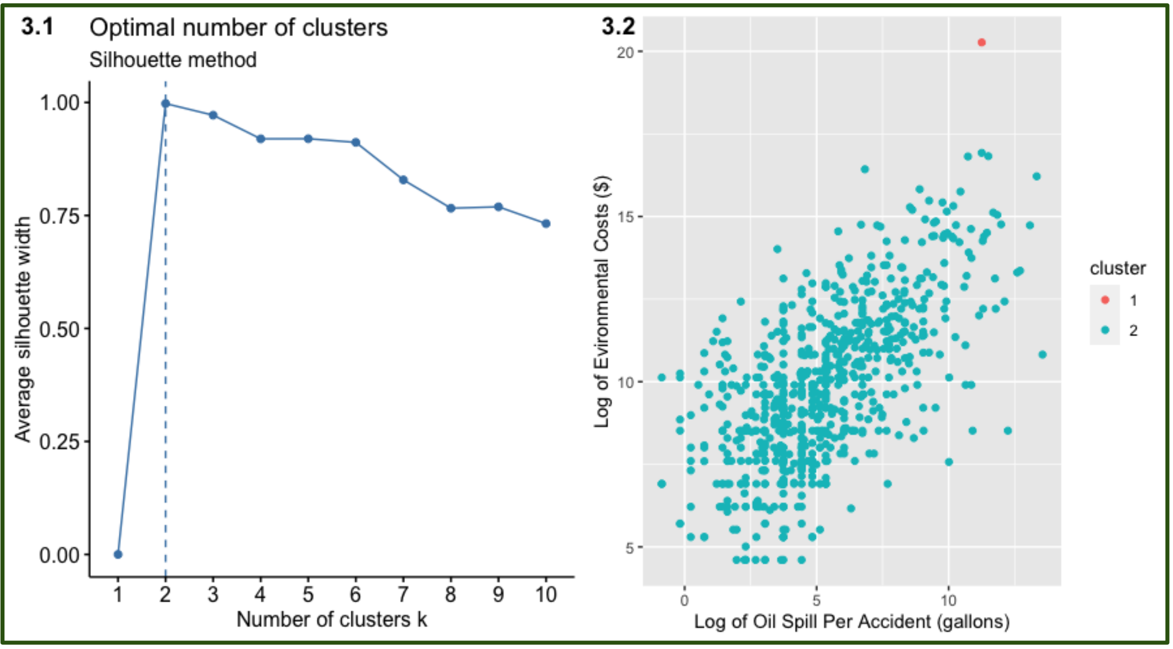
2.3 Distribution of the dependent and independent variables

The dependent variable (environmental cost) is extremely right skewed (skewness of 52) and ranges from $0 to $635,000,000. To visualize skewed data like this, it helps to consider the log of the values (Appendix 6.3). Appendix 6.3 illustrates that in the most cases, environmental costs are quite low, but that they can also be extremely high for some accidents. A skewed dependent variable (especially one as severely skewed as this one) is undesirable because it will reduce the accuracy of linear regression models. Therefore, if a linear regression were used to try to detect outliers in the dataset, the accuracy of the results would be negatively impacted by the skewness of environmental costs. Before addressing the skewness of environmental costs, let's take a deeper look at its values in the dataset.

[3] The original dataset from the accident reports has a substantial number of missing values, so much so that some columns (variables) had to be removed altogether because there was an insufficient number of observations for them. Please see the code file, "FinalProject.R" for a step-by-step description of the data pre-processing.

Another way to visualize the environmental costs is using clusters (See Figure 3). Using k-means clustering to group the accidents by non-zero environmental cost and non-zero loss of oil (in gallons) reveals two characteristics of the dataset. One, there is one accident (the sole accident in cluster 1) with environmental costs that are much higher than any other accident. This accident was likely contributing significantly to the positive skewness observed earlier. Two, the environmental costs of an accident vary greatly, even for the same level of oil spilled. For example, in Figure 3.2, the environmental costs of an accident range from approximately $500 (log(500) ≈ 6) to $1,000,000 (log(1000000) ≈ 14) for accidents where about 150 (log(150) ≈ 5) gallons of oil was spilt.

Figure 3



Due to the skewness and variability of environmental costs, rather than predict the exact environmental cost of an accident, one can classify accidents as minor or severe, depending on if the cost is above or below a certain value. The average environmental cost of the accidents is $364,075. Accidents will therefore be classified as severe if the environmental cost is equal to or above $364,075, and minor if it is lower. This transformation will make our dependent variable into a binary variable with a value of 1 if the environmental cost is severe, and 0 if it is minor.

Before selecting and tuning a predictive model, it is prudent to analyse the distribution of the fourteen variables in the dataset that are the possible predictors for the model. Four of these variables are numeric, six are binary, and four are categorical. The numeric variables are the year of the accident and the quantities of oil released unintentionally, intentionally,[4] and recovered during the accident. Unlike the year variable, which is roughly normally distributed (Figure 1), these other numeric variables are extremely right skewed (see Appendix 6.4). This indicates that in most accidents, there is not much oil released (unintentionally or intentionally) or recovered.

Pie charts of the six binary variables revealed that 99% of the accidents were onshore, 97% of them did not lead to a fire, and 99% of them did not lead to an explosion. The values for the remainder of the binary variables are more evenly distributed. See Appendix 6.5 for a visualization of how the other three binary variables are distributed. Finally, the four categorical variables (state, operator name, primary cause, and secondary cause) are right skewed as well. The right skewness of these variables indicates that most of the accidents happen to pipelines that are owned by a minority of the pipeline operators, and that the same primary and secondary causes are responsible for the accidents. For example, 19% of the accidents happened to pipelines owned by the same three operators: Enterprise Crude Pipeline LLC, Sunoco Pipeline L.P., and Plains Pipeline. Appendix 6.6 visualizes the top primary and secondary causes of accidents.

3 Model Selection and Methodology

3.1 The Relationships Between Independent Variables

The first step of selecting the independent variables for the model is evaluating how these variables relate to each other. Recall that each of these variables represent a characteristic of the pipelines or the pipeline accident, so intuitively, it is likely that they are related. For instance, underground pipelines could be less likely to explode, which would result in a negative correlation between pipelines being underground and them exploding during an accident. To evaluate the relationships amongst the binary and numeric variables in the dataset, a Pearson correlation test was run. The results of this tests are visualized in a heat map (Appendix 6.7). Interestingly, the test indicates that there is *not* a significant level of correlation between any of the variables (all below 0.5); however, the correlation between the oil spill igniting and exploding is 0.39, which is relatively high. Therefore, it could be detrimental to include both these variables in the model. To double check for correlation amongst the non-categorical independent variables, a logistic regression was run with severe environmental costs as the binary response variable, and the ten numeric and binary variables as the predictors. Then a variance inflation factor (VIF) test was used to check for collinearity amongst the predictors. No pair of predictors exhibited collinearity, not even the variables representing if the

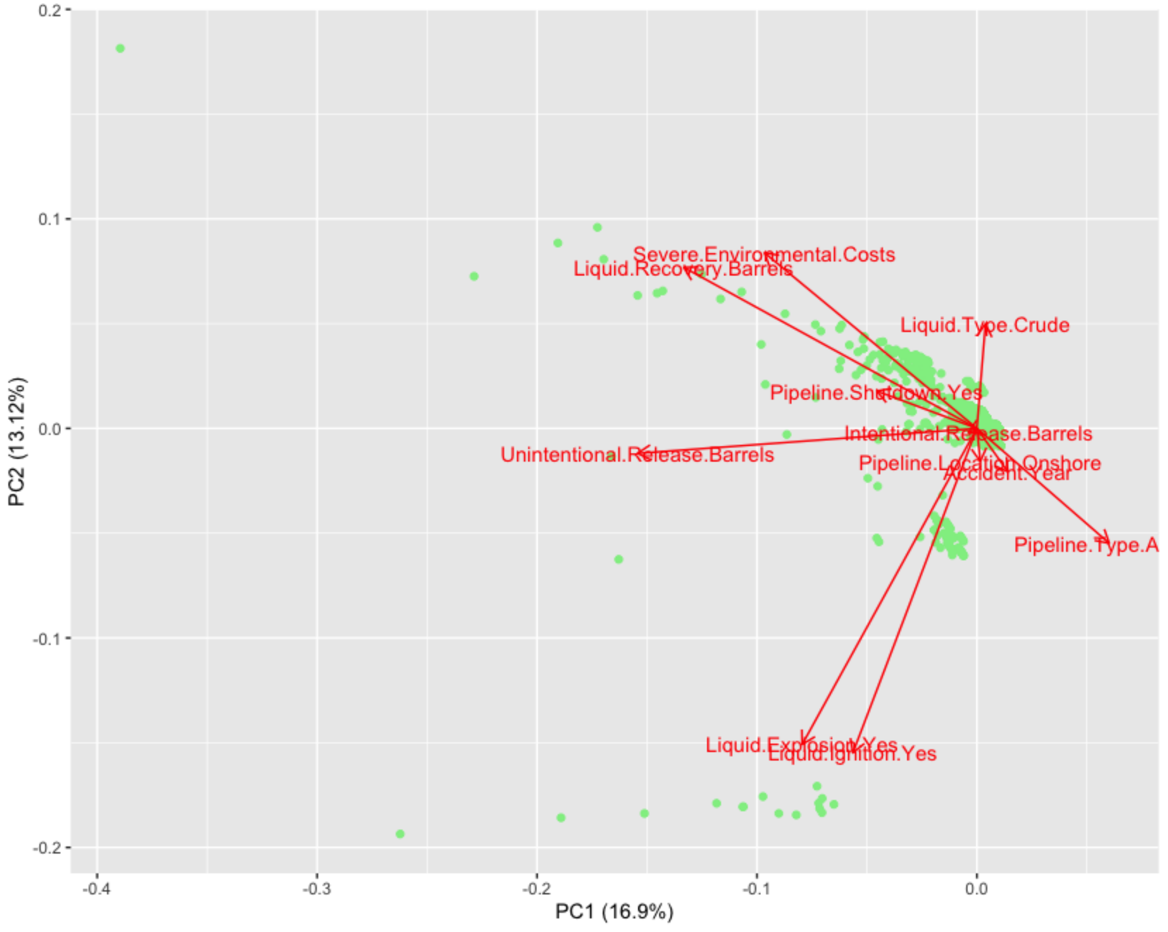
[4] From the description of the dataset, it is unclear if intentionally released oil refers to oil that is released to test cleanup technology's, vandalism of pipelines, or companies intentionally releasing petroleum products from a pipeline.

oil ignited and if it exploded, which only had a VIF score of 1.19 and 1.30 respectively (the threshold for collinearity is a VIF score above 4). At this point, there is no evidence of significant correlation or collinearity that would rule out certain predictor variables from being included in the model.

3.2 The Relationship Between The Dependent Variable and The Independent Variables

The next step is investigating the relationships between the numeric and binary variables and environmental costs. Understanding these relationships will help determine which characteristics of a pipeline, or a pipeline accident, affect the severity of the environmental damage of an accident. For example, pipelines that run above ground may have more environmentally damaging accidents. A principal component analysis (PCA) of our dataset elucidates the relationships between the variables (Figure 4).

Figure 4: Principal Component Analysis



In Figure 4, the green points represent the pipeline accidents, and the red arrows correspond to the non-categorical variables in the dataset. When the arrows are close together and pointing in the same direction, this indicates that the variables they represent are correlated.

The variables that are correlated with severe environmental costs are going to be the most useful predictors in a model. Thus, Figure 4 indicates that the amount of oil recovered after an accident and the pipeline being shut down by the accident are correlated with the accident having severe environmental costs associated with it. Additionally, arrows that point in the opposite direction of the severe environmental costs arrow represent useful predictor variables because these variables are negatively correlated with environmental costs. Thus, Figure 4 indicates that accidents for underground and offshore pipelines are correlated with severe environmental costs since the above ground and onshore binary variables are negatively correlated with the outcome variable. Finally, when an arrow is orthogonal, or close to orthogonal, to the environmental costs arrow, this indicates that the variable it represents is not correlated with severe environmental costs. Based on Figure 4, the ignition or explosion of oil during an accident are not correlated with severe environmental costs. Table 1 summarizes the evidence from the PCA.

Table 1: Evidence from The Principal Component Analysis

Useful predictor variables	Potentially useful predictor variables	Weak predictor variables
Liquid.Recovery.Barrels	Unintentional.Release.Barrels	Liquid.Explosion.Yes
Pipeline.Shutdown.Yes	Liquid.Type.Crude	Liquid.Ignition.Yes
Pipeline.Type.Above		Accident.Year
Pipeline.Location.Onshore		Intentional.Release.Barrels

The insights from Figure 4 are a helpful starting point for selecting the independent variables for the model, but they are not definitive. This is because Figure 4 only includes the first two principal components, out of the 11 that were created during the PCA, and a percentage-of-variance-explained plot (Appendix 6.8) indicates that the first two components only capture 30% of the variability between pipeline accidents. Intuitively, this means that with the first two principal components alone, one can only identify 30% of the variability across the accident's dataset. Therefore, further investigation of the relationships between variables is needed before the final predictors of the model can be selected.

3.3 Using A Random Forest to Identify Predictors

A random forest model with severe environmental costs as the outcome variable, and all the other variables (except for the name of the pipeline operator), provides additional evidence about the usefulness of each predictor. The operator name variable has too many categories to be used in the random forest model,[5] but other categorical variables can be included (unlike during the PCA). The results of the random forest model indicate that the most useful predictors are the amount of oil unintentionally released, the amount of oil recovered, if the oil is crude or not, the state the accident happened in, and if the pipeline was shutdown. This result confirms some of the findings of the PCA (Figure 4 and Table 1). A summary of the results of the random forest model are shown by the variable importance plot in Appendix 6.9. In the next section, the evidence from sections 3.2 and 3.3 will be used to construct the final model.

3.4 Final Model Selection: Boosted Forest Model

The variables that were identified by the PCA as useful and shown by the random forest model to decrease mean squared error by more than 5% were selected to be the predictors in the final model. The final model is a boosted random forest because of the accuracy that it provides, but also because the algorithm is able to handle the operator name variable as a predictor. The later reason was particularly important because the name of the operator turned out to be the most useful predictor (see Appendix 6.10 for the relative importance of each predictor in the final model). The final model to predict if a pipeline accident will be severe or minor is the following:

```
boosted=gbm(Severe.Environmental.Costs~
        Operator.Name + Accident.State + Cause.Subcategory +
        Unintentional.Release.Barrels + Liquid.Recovery.Barrels + Pipeline.Shutdown.Yes +
        Pipeline.Type.Above + Liquid.Type.Crude,
      data=spills_final, distribution="bernoulli", n.trees=10000,  interaction.depth=6)
```

4 Results

The final boosted forest model predicts the severity of pipeline accidents, but it also provides insights as to which characteristic of an accident are correlated with severe environmental damage. The model reveals that the name of the operator, the cause of the accident, the quantity of oil unintentionally released, the quantity of oil recovered immediately after the accident, and the state in which the accident took place are the most influential predictors of the severity of environmental damage.

The model's prediction accuracy was tested by training it on 85% of the accidents in the dataset and then randomly sampling 15% of the accidents 100 times and predicting the severity

[5] The operator name variable has 227 levels. The RandomForest() function in R has a limit of 53, for the number of levels a categorical variable can have. This limit is necessary because of unreasonable (or even infeasible) amount of time it would take the random forest algorithm to run if there was a variable with more than 53 levels. Consider the operator name variable with 227 levels. If it were in the random forest model, the random forest would have to perform $2^{227-2} = 2^{225}$ possible divisions of the operator variable to find the best option, which means it would have to calculate and evaluate 5.3919893e+67 results! If it took the computer 0.00001 seconds per result, it would take 5.3919893e+62 seconds… as of now, the Earth has only existed for $1.433e^{17}$ seconds. There is a good reason for this limit!

of the accidents each time. The average error from this test was approximately 0.08 or 8%, indicating that the severity of a pipeline accident can be predicted with a high degree of accuracy by the eight predictor variables in the model. The final model and the random forest model used in section 3.3 were also tested with a dataset that did not include observations identified as outliers. The results for both models were not significantly different when the dataset without outliers was used. Therefore, given that the dataset is quite small to begin with (2777 observations), the decision was made to use the complete dataset, rather than the one without outliers, for testing models. The final section of this report will take a closer look at each of the predictors in the final model and discuss the implications and limitations of this analysis.

5 Lessons and Conclusion

The goal of this analysis was not to predict if a pipeline accident will happen, but rather to predict how severe an accident's environmental damage will be. The final model accurately predicts if an accident will cause severe or minor environmental damage, and indicates that there are four primary drivers of environmental damage severity: the operator, the cause of the accident, the quantity of oil spilt and recovered, and the state it took place in. Therefore, to mitigate the threat that pipeline accidents pose to the environment, it is recommended that governments focus on regulating on these aspects of pipelines. For instance, operators could be taxed at a higher rate if they have a history of environmentally damaging accidents.

The model has a few short comings that should be mentioned. Some of the variables used to predict the outcome, such as the cause of the accident or the quantities of oil released and recovered during the accident, are not known until a while *after* the accident occurs. This means that this model cannot be used to predict the immediate environmental damage of an accident. Nevertheless, it is still useful because accounting for environmental damage after an accident is a very tedious and complicated endeavour.[6,7] This model could be used as a preliminary tool to estimate the environmental damage caused by an accident before experts have the time to fully account for all the damage associated with the accident.

Prediction is the primary use of the model, but it can also be used to learn about which characteristics of a pipeline and an accident affect the severity of environmental. The model

[6] Goldstein, M., & Ritterling, J. (2001). *A Practical Guide to Estimating Cleanup Costs*. Digital Commons. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1027&context=usepapapers

[7] INTERNATIONAL ATOMIC ENERGY AGENCY. (2019). *Developing Cost Estimates for Environmental Remediation Projects*. https://www-pub.iaea.org/MTCD/Publications/PDF/P1857_web.pdf.

indicates that the primary determinants of environmental damage are the company operating the pipeline, the cause of the accident, and the quantity of oil spilt during the accident. The latter two are not surprising, it makes sense that the cause of an accident will influence how damaging the accident is, and similarly, it makes sense that the amount of oil spilt will impact environmental damage. On the other hand, the significant effect of operators on the environmental damage caused by an accident is interesting. This relationship could be caused by the fact certain operators are more prone to experiencing accidents because of how they build, maintain, or monitor their pipelines. Alternatively, the connection between environmental damage and the operating company could also exist because certain companies spend more on environmental remediation costs after an accident. This would mean that the model is predicting whether a lot of money will be spent on the environmental clean-up, but not necessarily if the damage the money is spent on is significant. This is a drawback of this model. In the absence of detailed data about the damage done to the environment by the accidents, the model uses environmental remediation costs as a proxy for environmental damage. Consequently, it is limited by the strength of the relationship between the money that is spent repairing the environment and the actual damage the environment endures. This limitation could be resolved by incorporating environmental data. For instance, data about the toxicity of water near a pipeline accident could be used as a new outcome variable that measures the impact an accident has on the environment. The model could also be improved by incorporating more recent data, as it becomes available.

Environmental protection and preservation were the fundamental motivations for this report. There is really nothing environmentally friendly about pipelines, even if they do not have an accident, they are still transporting fossil fuels. So obviously, the best option of for our environment would be for us to quit using fossil fuels and thus pipelines altogether, but unfortunately, we do not seem to be headed there anytime soon. In the meantime, one of the things that we *can* do is leverage technology to mitigate the damage we do to our planet every day. This project, at the very least, is a demonstration of this. It demonstrates how machine learning techniques and statistical analysis can help us describe, predict, and ultimately prevent environmental harm.

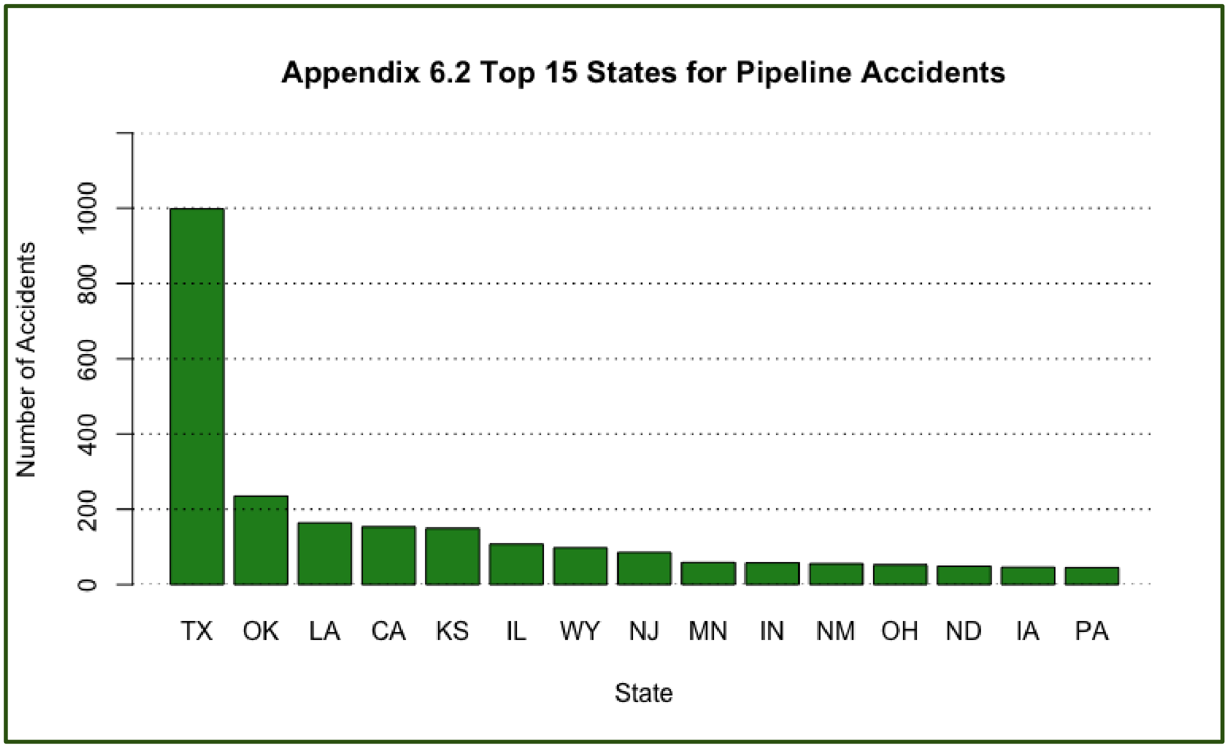
6 APPENDICES

APPENDIX 6.1 – DATA DICTIONARY

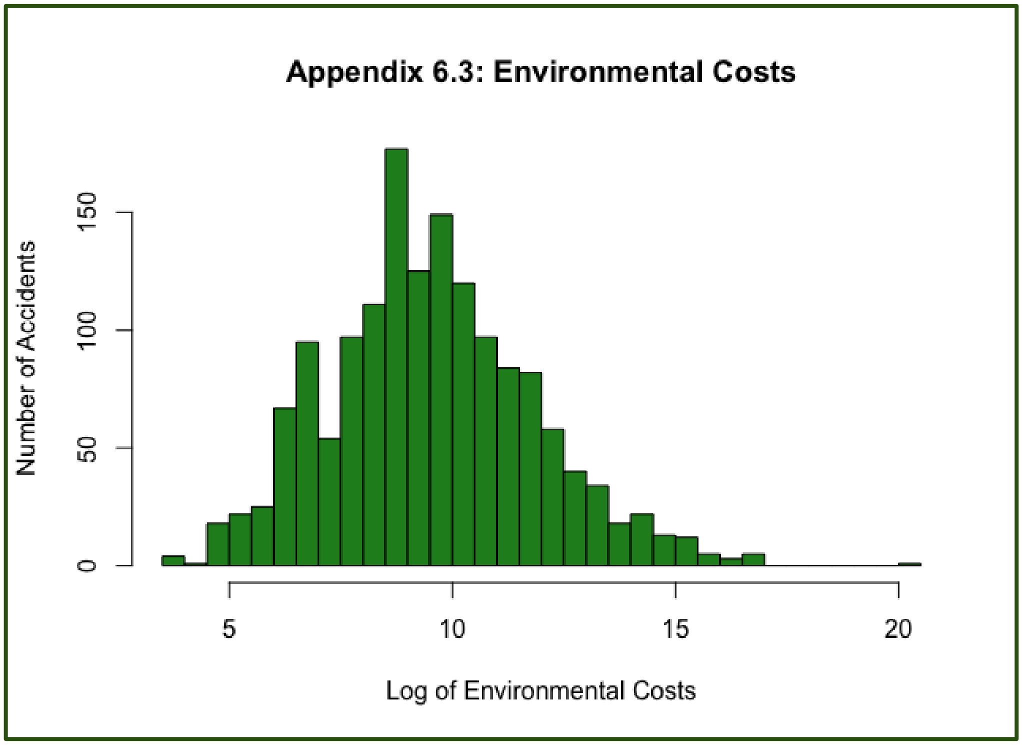
Dependent (Outcome) Variable
Severe.Environmental.Costs: 1 if environmental damage from accident is severe, and 0 if it is minor.

Independent (Predictor) Variables
Accident.Year: Year of the pipeline accident (2010-2017)
Operator.Name: Name of the company that operates the pipeline
Accident.State: State the accident happened in
Cause.Category: Primary cause of the accident
Cause.Subcategory: Secondary cause of the accident
Unintentional.Release.Barrels: Quantity of oil (in barrels) unintentionally spilled during the accident.
Intentional.Release.Barrels: Quantity of oil (in barrels) intentionally spilled during the accident.
Liquid.Recovery.Barrels: Quantity of oil (in barrels) recovered after the accident.
Liquid.Ignition.Yes: Did the oil ignite during the accident? (1 for yes, 0 for no)
Liquid.Explosion.Yes: Did the oil explode during the accident? (1 for yes, 0 for no)
Pipeline.Shutdown.Yes: Did the accident shutdown the pipeline? (1 for yes, 0 for no)
Pipeline.Location.Onshore: Is the pipeline onshore? (1 for yes, 0 for no)
Pipeline.Type.Above: Is the pipeline above ground? (1 for yes, 0 for no)
Liquid.Type.Crude: Is the pipeline carrying crude oil? (1 for yes, 0 for no)

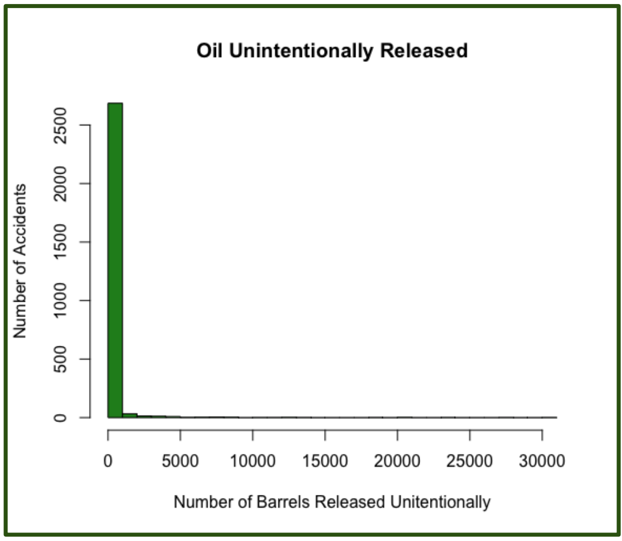
Appendix 6.2 – Top 15 States for Pipeline Accidents

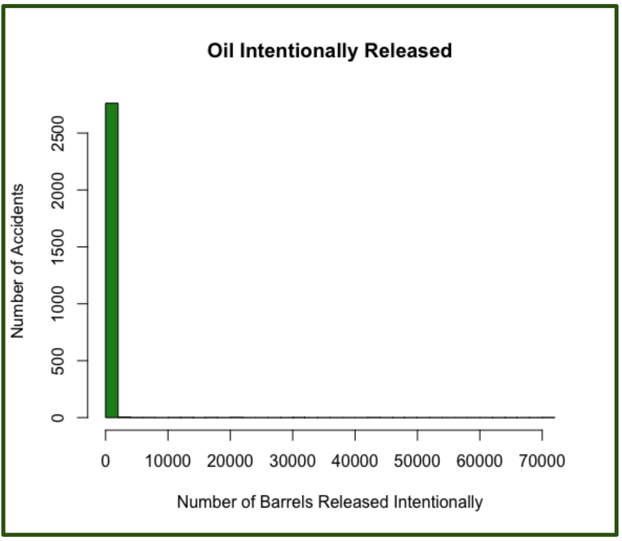


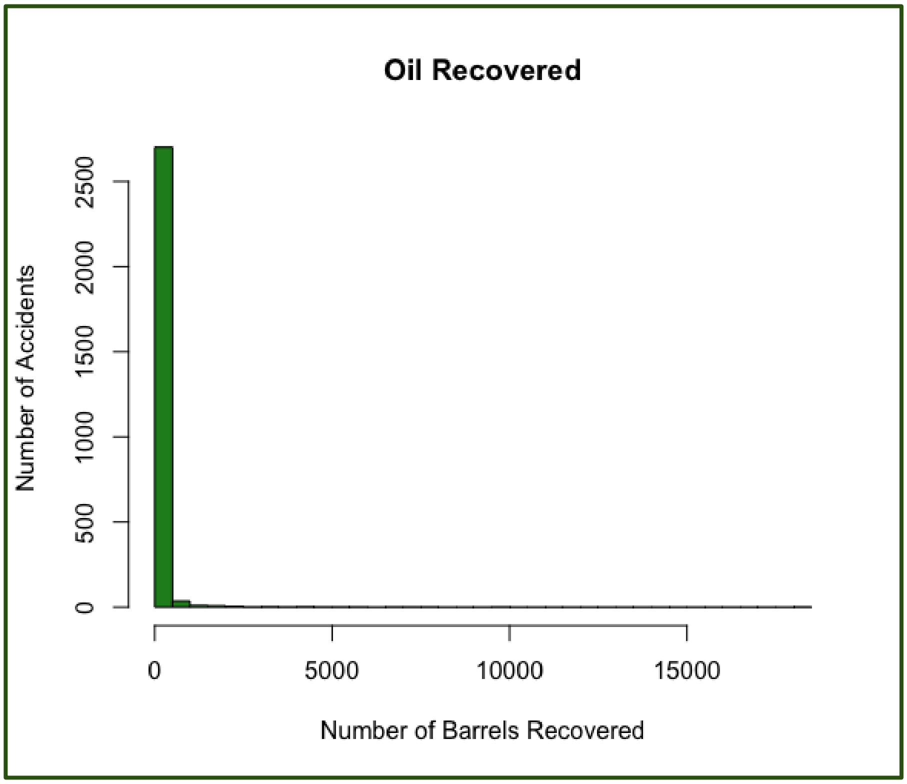
Appendix 6.3 – Environmental Cost of Pipeline Accidents



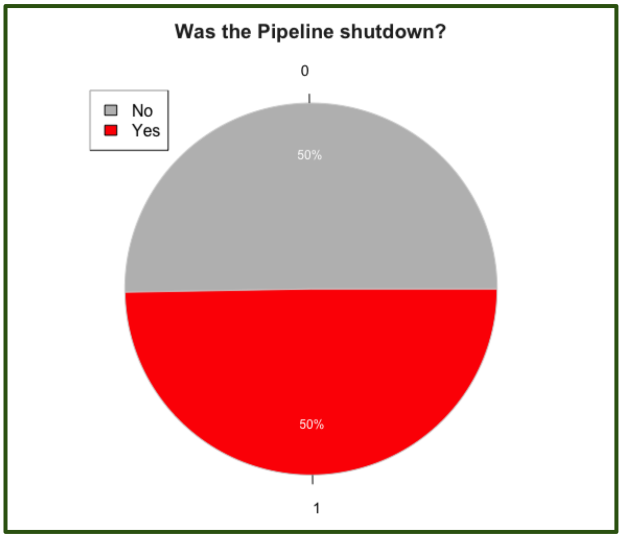
Appendix 6.4 – Distribution of Numeric Variables

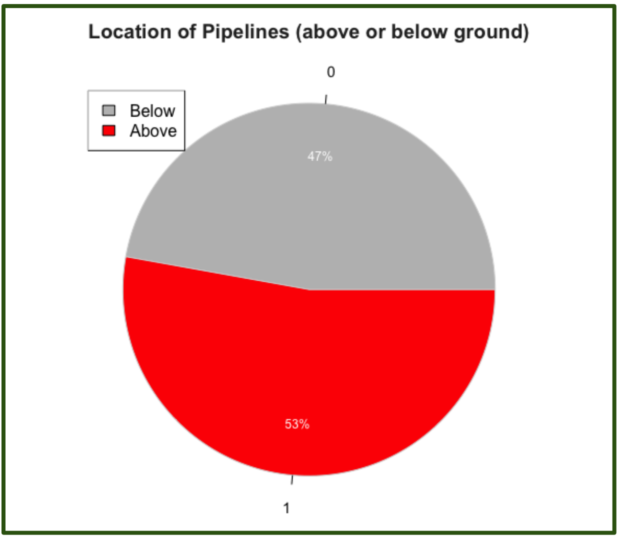


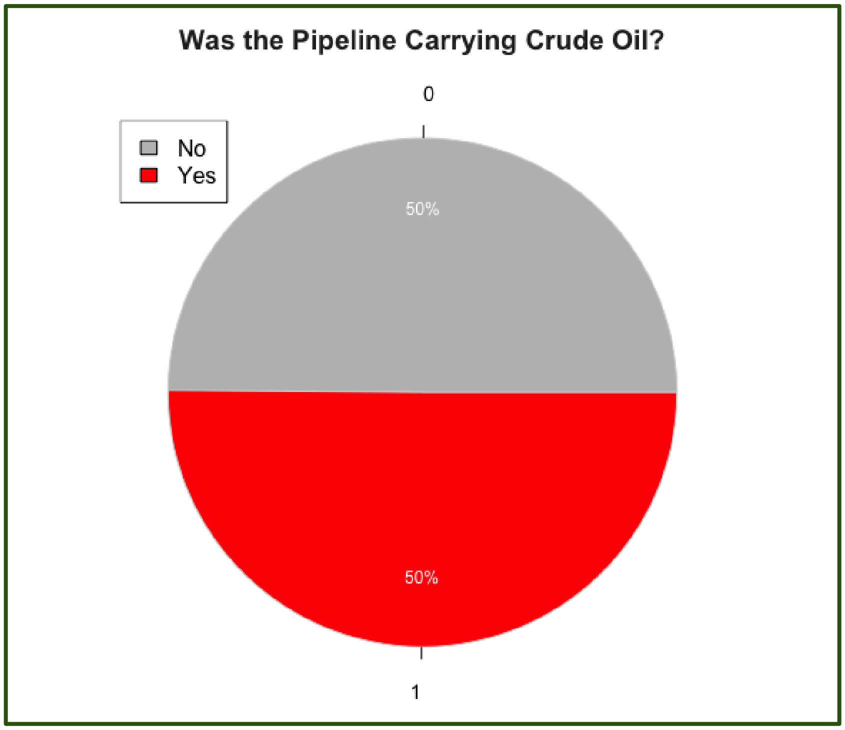




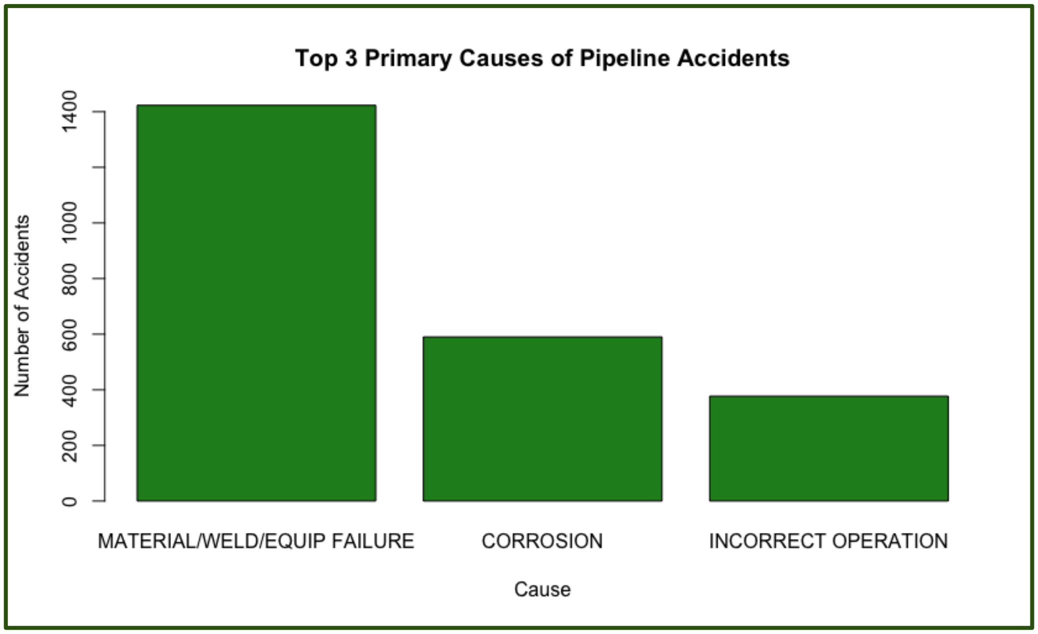
Appendix 6.5 – Distribution of Binary Variables

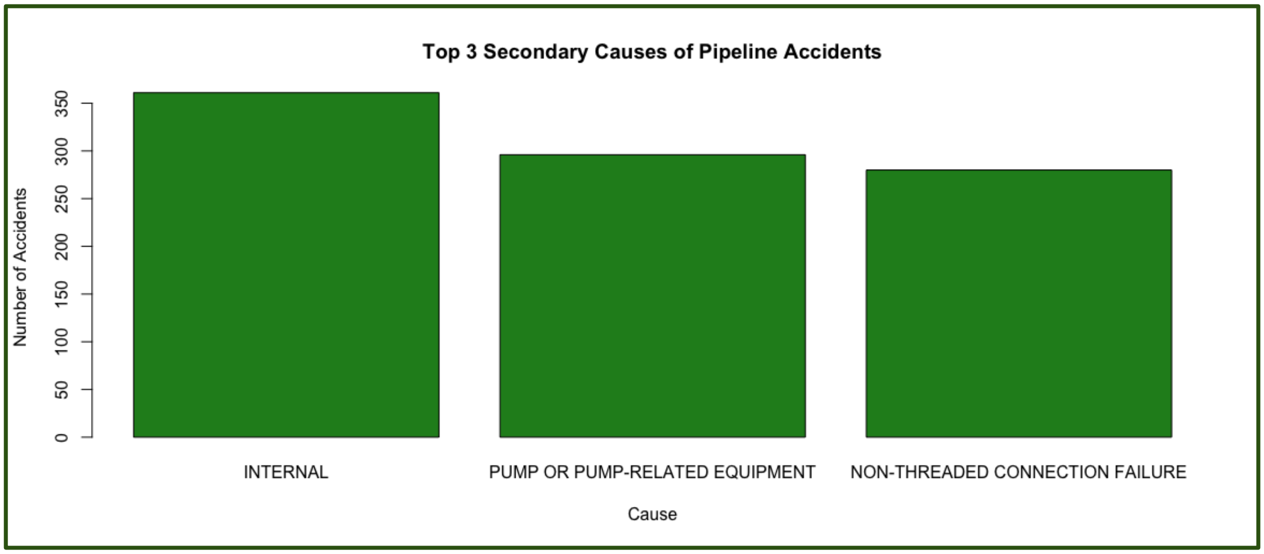




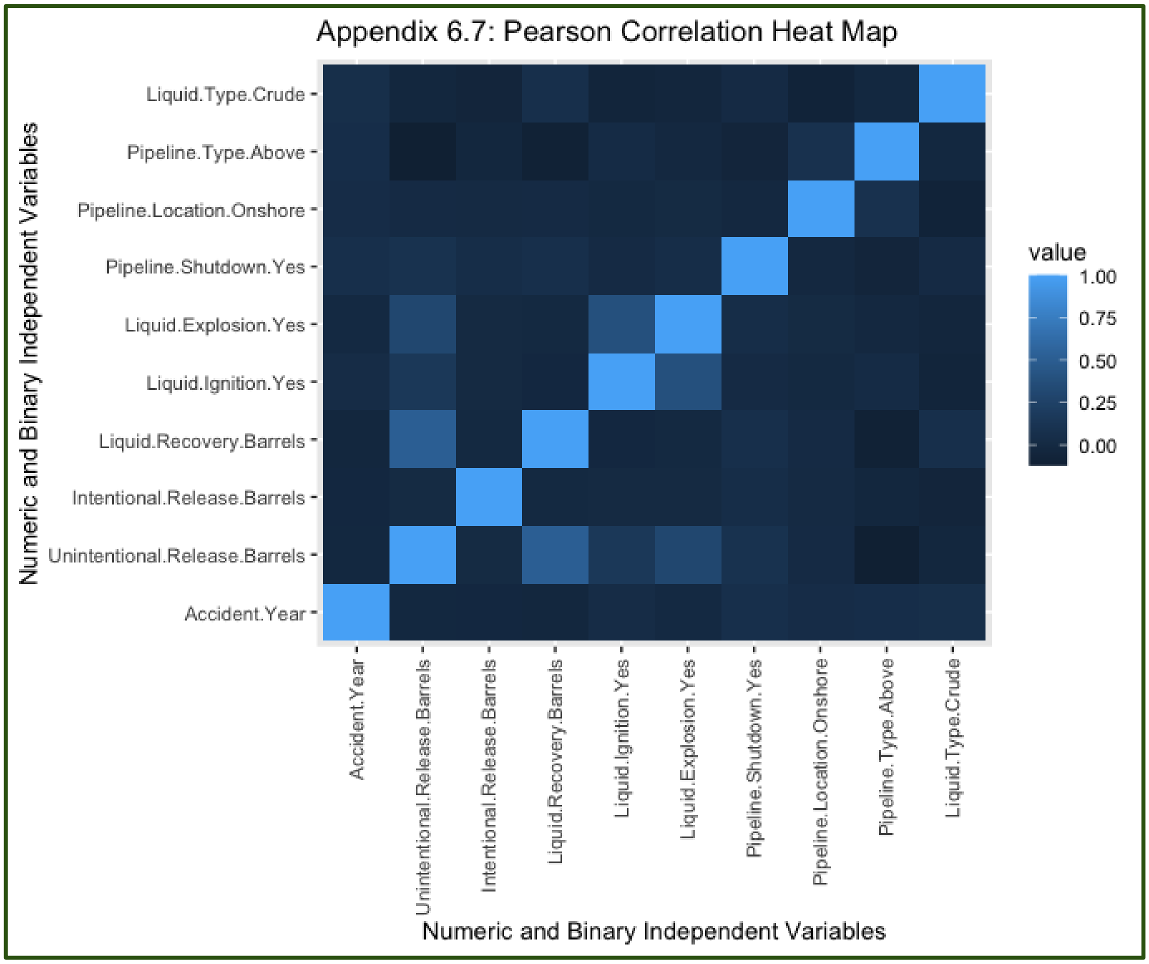


APPENDIX 6.6 – DISTRIBUTION OF CATEGORICAL VARIABLES

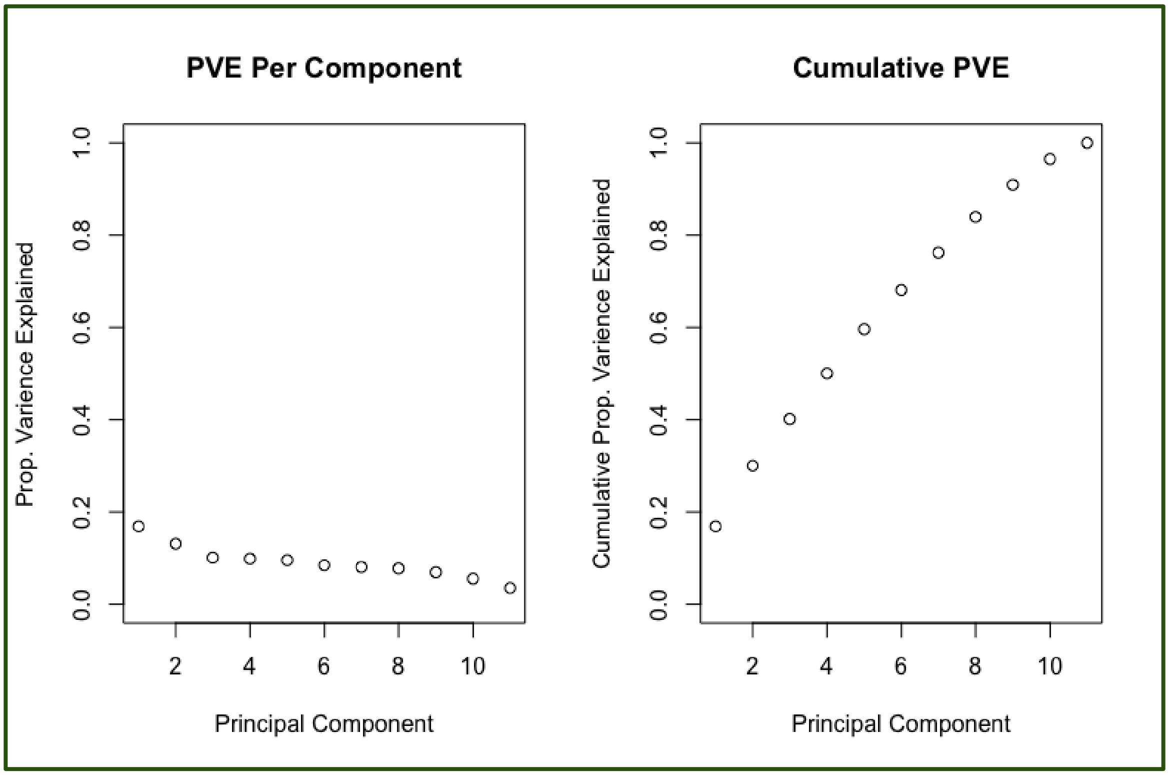




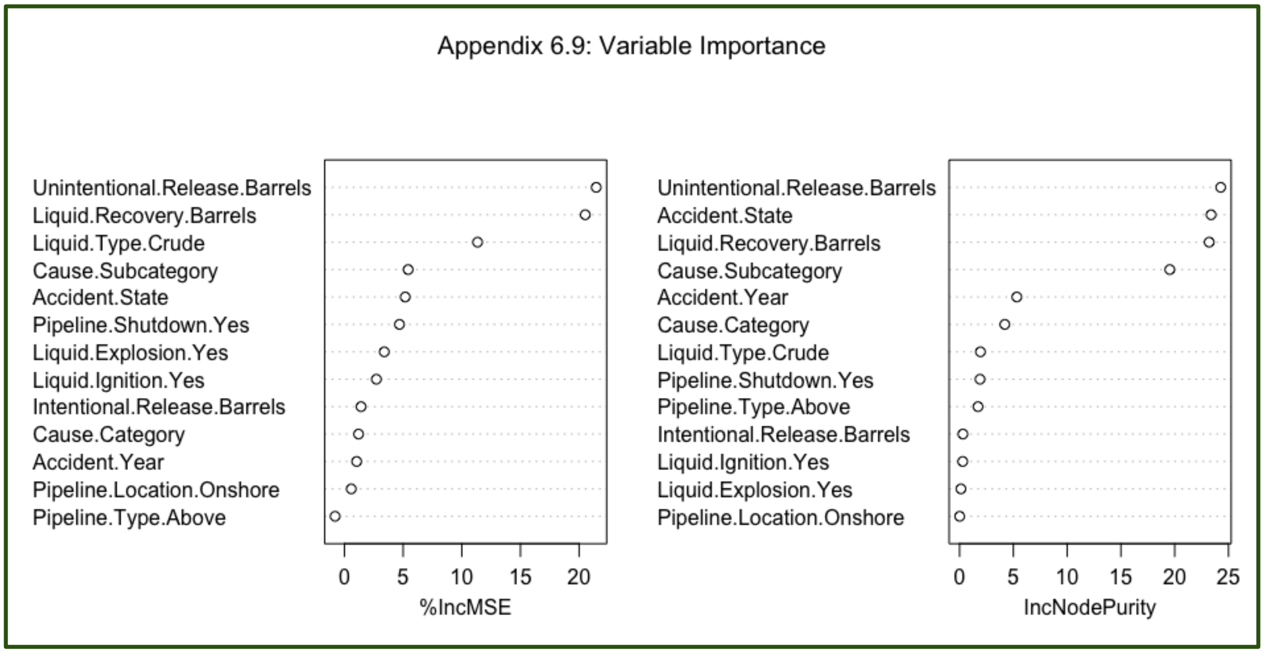
Appendix 6.7 - Pearson Correlation Matrix Heat Map



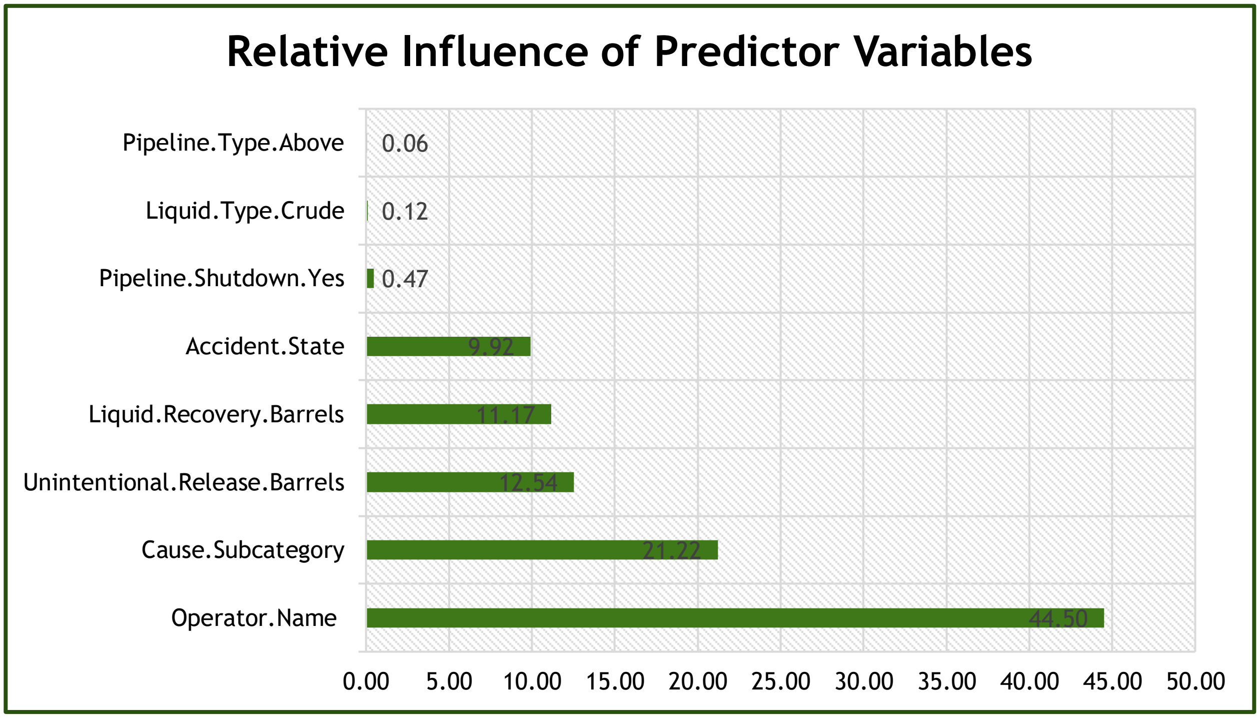
Appendix 6.8 – Principal Component Analysis Percentage-of-Variance-Explained (PVE) Plot



Appendix 6.9 – Random Forest Variable Importance Plot



APPENDIX 6.10 – RELATIVE INFLUENCE OF EACH PREDICTOR IN THE FINAL BOOSTED FOREST MODEL



7 CODE

Please see R script that was included with the submission of this report, entitled: “FinalProject.R”

The End. Thanks for reading!

Michael Church Carson (260683849)