# CHAPTER FOUR: Building Predictive Model by using Machine Learning Algorithms

* Introduction
* Data Description, and data Preparation
* Choosing Machine Learning Algorithms
* Modeling Methodology
* Processing The Predictive Model
* Model Testing, and Validation
* Discussion of Results, and Future work.

## **Introduction**

The aim of this chapter is to develop a predictive model by using machine learning algorithms to predict the gas pipe damages. This chapter consists of Introduction, Data Description, and data Preparation, Modeling Methodology, Machine Learning Algorithms used in the analysis, Processing The Predictive Model, Model Testing, and Validation, Discussion of Results, and Future work. In addition, this chapter will answer the following research question:

1. Design risk effective model to predict future risk in UG gas pipe damages by using Machine Learning Algorithms by studying the past underground gas line damages in urban congested cities?

## **Data Description, and data Preparation**

**THE BODY WILL BE ADDED LATER**

## **Choosing Machine Learning Algorithms**

* Logistic Regression Analysis.

Logistic Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (Target) which is Damage, OR Undamaged and independent variable (s) (Predictor) data attributes, Time, Days, weeks, Months, Years, Cities, Counties, D10, D20, D30. This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, relationship between Damage and undamaged UG gas pipes will be determined through the attributes of the predictive model. More specifically, the regression analysis will go through each point of attribute (predictor) and try to fit most of the points through patterns. In addition, regression analysis is an important tool for modeling and analyzing data. In our study, regression analysis fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. More details will be explained later, how the logistic regression algorithm works within our data.

Why we employed Regression Analysis in our study Model.

1. Regression Analysis indicate the significant relationships between dependent variable (Target (YES/NO)) and (Predictors (data Attributes)) independent variable.
2. Regression Analyses indicate the strength of impact of multiple independent variables (data attributes) on a dependent variable (Damage/ Not).

* Bayesian Analysis.

Bayesian Algorithm work by an assumption of independence among predictors (Data Attributes). In more details, a Naive Bayes classifier assumes that the presence of a particular attribute in a class is unrelated to the presence of any other Attributes.

Why we employed Bayesian Analysis in our study Model.

1. It is easy and fast to predict class of test data set. It performs well in case of categorical input variables compared to numerical variable(s). Most of the data used in the study has label attributes such as binary, categorical or nominal. Is classified as categorical variables (1, 0).

* [k-nearest neighbor’s algorithm](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm) (K-NN).

K nearest neighbors is a simple algorithm that stores all available cases (Predictors) and classifies new cases based on a similarity measure (e.g., distance functions).

Why we employed KNN Analysis in our study Model.

1. KNN was selected to be used in this study because it works as pattern recognition though specific distance for each neighborhood, then it classifies the data based ion the distances between the generated neighborhoods.

* Random Forest.

Random forest builds multiple decision trees for the data attributes Time, Week days, Months, and other attributes and merges them together to get a more accurate and stable prediction. More specifically, it creates a forest of data attributes, and makes random selection.

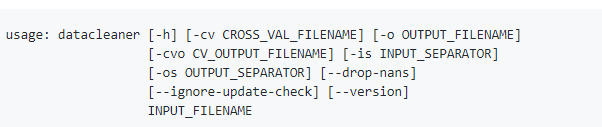
Why we employed Random Forest Analysis in our study Model.

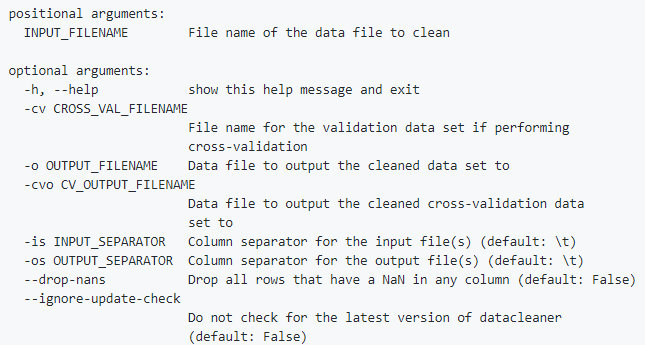
1. Random forest can be used for both regression and classification tasks and that it’s easy to view the relative importance it assigns to the input Attribute.

## **Predictive Model Methodology**

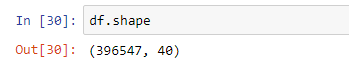
Now we understand the machine-learning problem want to solve for. Which is predicting gas pipe damages for future UG gas pipe operations? The next step is to build a model which is to employ data science methodologies like Logistic regression, Random Forest, KNN, Bayesian. Looking at the historical data we have, we want to produce a model that estimates a particular variable specific. Which is (YES/NO) damages or undamaged. The following steps explain the preliminary steps to input data into Python (Anaconda)

* Data Cleaning as following.





* Total data used 396,547 ( including undamaged & damages).



* Total attributes/features = 40 as a start , then developed in python into 752 attributes, cities were put in the columns and values (1, or 0) was assigned based on the ticket, damaged or not.
* Data cleaning was performed in Excel, and Python Anaconda.
* The data split was 80% Training, and 20% Testing



Which means; Data in Training = 0.8 \*396,547 = 317,237

Data in Testing = 0.2\* 396,547= 79,309

* Four Algorithms where used in the exploratory analysis (Logistic Regression, Random Forest, and K nearest neighbors, Bayesian).
* Micro was selected from (Micro, Macro, Binary).

The following chart explain the methodology of the Model. The process starts by preparing the data, then inputting data into Anaconda, then cleaning the data again, then splitting the data, then running the algorithms in 80 % training. Then we run remaining 20% testing data into testing and test the model. The following step is to select a model based on the testing matrixes such as confusion matrix, precession, recall. then run the confusion matrix, precession, recall.

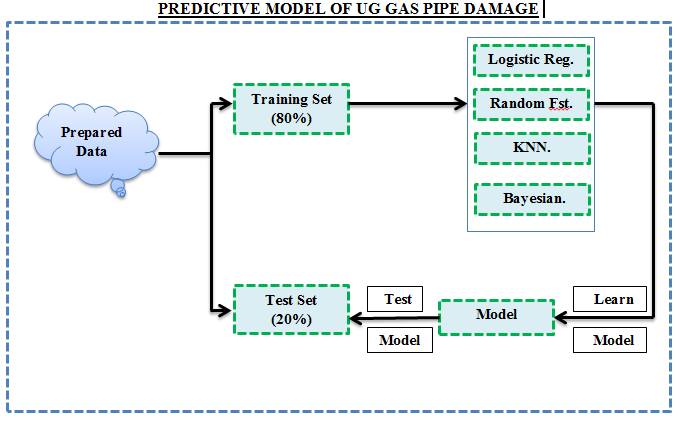


Figure 1: Predictive Model for Underground Gas Pipe Damages.

## **Processing the Predictive Model (Training The Model)**

The process of training a predictive model involves providing PM algorithm (that is, the learning algorithm) with training data to learn from. The term PM model refers to the model artifact that is created by the training process. The provided training data to the algorithm contains the correct answers, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an PM model that captures these patterns. We will use the model to get predictions on new data for which you do not know the target. In our study we provided the algorithm with 40 attributes called predictors, and the Target which is YES or No, damage or undamaged see figure Figures 2, 3 .

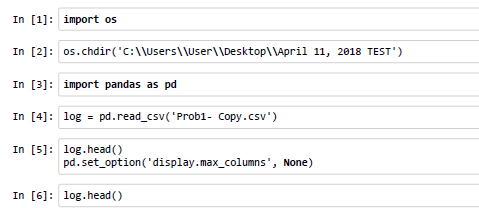


Figure 2 Importing Data & Libraries

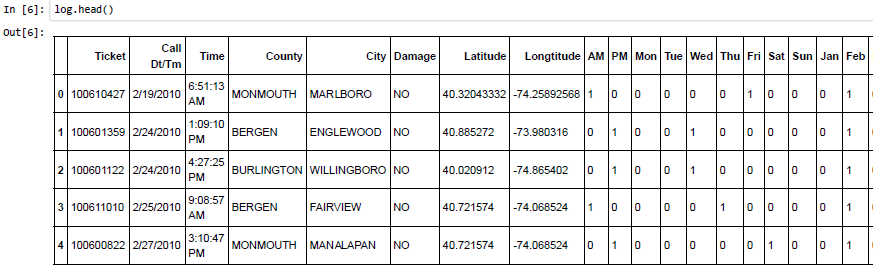


Figure 3 Data Attributes

This step is viewing the attributes type, and checking the attributes' normal distribution, for the predictors to give good results, all of the data attributes needed to be normally distributed. Thus, standard deviation test was performed on some of attribute of the data. The results were some of the attributes were not normally distributed, Python was used for that test see figure 4.

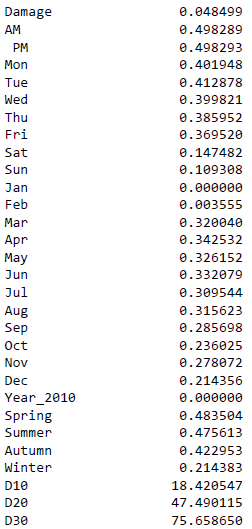


Figure 4 figure shows the standard deviation of all attributes

As can be seen in figure 4, damages within 10 miles diameter, 20 miles diameter, and 30 miles diameter have the values 18.420547, 47.490115, and 75.658650. These values will have negative impact on the modal. Thus, all these three values were normalized for all data set. Standard deviation equation was used to normalize these values, see figures 5.

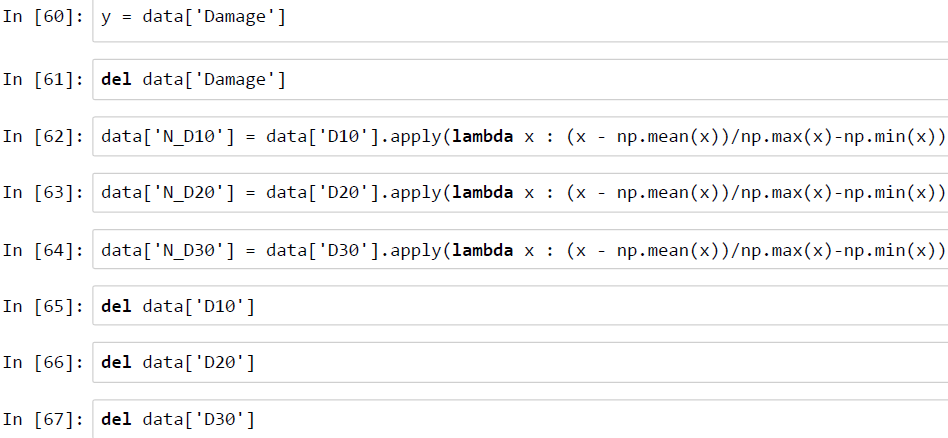


Figure 5 standard Deviation for D10, D20, D30

All Libraries, Methodologies, and needed tools were imported through Python Anaconda Library see figure 6.

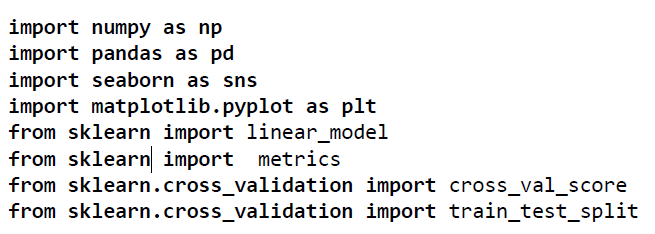


Figure 6 Importing Libraries to be used in the calculations

* Logistic Regression Algorithm

The followings are specific codes were used in Regression Algorithm and will be explained later in more details.

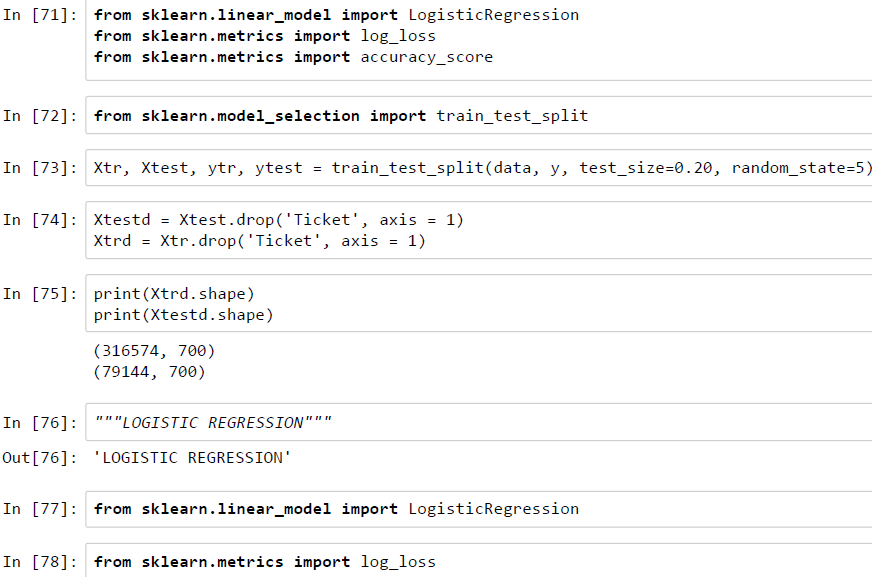


Figure 7: Logistic Regression Algorithm

* Random Forest Algorithm

The followings are specific codes were used in Random Forest Algorithm

and will be explained later in more details.

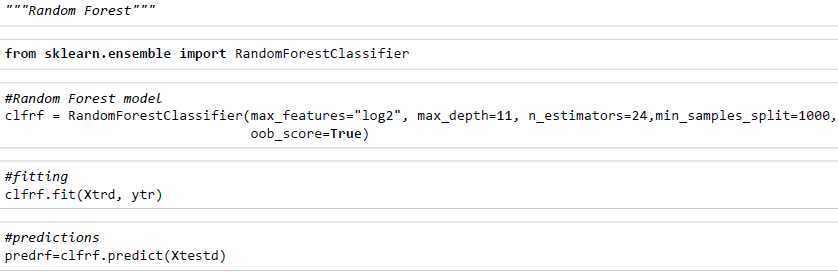


Figure 8 Random Forest Algorithm

* KNN

The followings are specific codes were used in KNN

and will be explained later in more details.

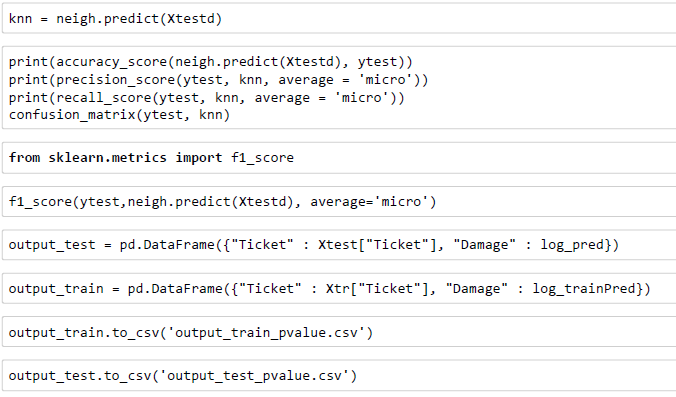
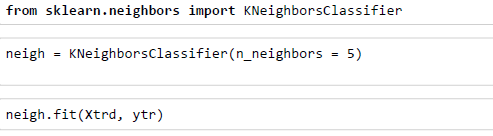


Figure 9 KNN Algorithm

Explaining the Model how it works. The data was divided to X, Y , where is X equal to all attributes, and Y equal to Damage/ Target Variables. The data split was 80 % of the data into training, 20% into testing. Moreover, 70% of the data is in X-train, Y-train. Then, the classifier was chosen to be either Logistic regression, KNN, Random Forest, Or Bayesian. Then the equation, CLF.fit(x-train, Y-train) finds the patterns in all the attributes with respect to Y- train, and save the patterns. Results, the algorithm, found the pattern in the training data. Next, the model uses the pattern found from the training data which maps all the attributes to target value which is damage (YES/NO). Using the pattern found above combined with another function called “PREDICT”. Equation Data-Predict = CCF. Predict ( X – test), Data – Predict = = ( Y – Test) Actual. Finally, we score the model or test the model by comparing Data-Predict Vs Y- Test.

## **Model Testing, and Validation**

## 20 % of the data was used in testing the predictive model, and the metrics were used to evaluate the model were Confusion Matrix, Recall, and Precision. These three Metrics were conducted on all of the outcomes of the algorithms including Logistic Regression, Random Forest, KNN, and Bayesian. Results, Logistic Regression Model gave the best result according to Confusion Matrix, Recall, precision. Sample was used for illustration purposes, see figure below. More details will be added later. In addition, comparison between confusion matrix for KNN, Bayesian, Logistic regression, and Random Forest will be added later.



Table 1show compared exporetd test data in excel



## **Discussion of Results, and Future work.**

The developed predictive model encompasses a variety of statistical techniques from modeling, machine learning, data mining and others that analyze UG gas pipeline historical facts to make predictions about future events. In our study, predictive models exploit patterns found in UG gas pipe historical data and selected attributes to identify the Target variable ( YES/NO) ( Damage/Undamaged) . The Produced Predictive Model capture relationships among many risk factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for agencies dealing with UG gas pipe digging process.

Even though, there are some limitation in this study, such as missing data, and not having some valuable data attributes such as pipe size, diameter, pipe materials type …etc. The developed model from the derived data attributes was able to predict more than 80 % of the future UG gas pipe damages.

In other hand, it’s advised as part for future work to collect more data attributes related the UG gas pipeline, such as pipe size, pipe materials, reason for damage, temperature of the gas pipe, and age of the gas pipe. Analysis can be further conducted after collecting these attributes to assess the impact on the predictive model, and accuracy of future prediction.

# CHAPTER FIVE: Identification of Dominant Factors Through Statistical Analysis

1. Introduction
2. Description of Data Sets
3. Data Pre-processing
4. Overview of Geo-code, Plotting, and Preliminary Exploratory Analysis of undamaged UG data
5. Preparing Data Features/ Attributes
6. Univaraite (Descriptive )Analysis
7. Bivariate Analysis
8. Discussion of Results, and Future Work.

## **Introduction**

**THE BODY OF INTRODCUCTION WILL BE ADDED LATER**

This chapter aims to identify the dominant risk factors affecting the underground gas pipe damages, an contributing to excavation damages to underground gas pipes. In addition, this chapter will help gain more understanding of how do these dominant factors relate to each other. Therefore, this chapter will answer the following research questions;

1. What are the dominating factors contributing to excavation damages to underground natural gas facilities?
2. How do these dominant factors relate to each other and form up a network of interacting factors triggering different excavation fates?

Python (Anaconda), and IBM SPSS Statistics were used to perform the analysis. In more details, Python was used to clean the data, repair the data, fix missing values, develop predictive model, and determine what the significant factors are. IBM SPSS Statistics was used to perform detailed statistics analysis to determine the significant factors, how these factors interact with each other, and how they contribute to the damage of the underground gas pipe damages.

## **Description of Data Sets**

The data used in this research includes excavation requests. Most of them did not result in damage to the underground facilities. Small portion of the excavation requests caused damage to the underground facilities. In addition, the data were received in the form of raw data: shown in Figure 10 and Figure 11. The analysis will be conducted on a period from years 2010 to mid-2013 and the data of year 2014 will be used to test and validate the results. The data set includes both damaged and undamaged UG gas pipes. All the received data was updated progressively as records were made available. In total, there were 2 Million records provided and used within analysis procedures.

The raw data was processed and organized in a way where rows represented individual incidents and whose columns represented provided attributes. For damaged date, 20 attributes were labeled for damaged data of UG gas pipes as follows:

1. Incident ID
2. Incident Date
3. Underground Facility Operator
4. Damage Address
5. Damage City
6. Damage Zip
7. Damage County
8. Excavator Name
9. Excavator Type
10. Excavator Address
11. Excavator City
12. Excavator StateExcavator Zip
13. Locate Ticket Number
14. Locate Ticket Type
15. Locate Performed
16. Damage Description
17. Pipe Material
18. Pipe Size
19. Cause of Damage

This is a large data set with approximately 2 Million entries. Unfortunately, some attributes within the data set were quite sparse because of the varying recording procedures among the four gas distributors. For the most part, important information was recorded; however, missing values are still noticeable within core data. Most of the variance in recorded values depends on who was responsible for updating records in master data sheets. While the recording system is most effective if all values are known, excavator names are not always known because work was

performed without requesting locate services and resulting damages were discovered after-the-fact. Another reoccurring error pertains to the accurate entry of data values within the master sheet. Frequently entries were recorded incorrectly into the wrong data cell; this caused a loss of information since the incorrectly recorded instances occupied cells of other required attributes. In summary, in this research we are using the excavation request data and those requests that resulted actual damages. For the excavation data set, we only have several attributes.

Figure 10: Received Raw Data Sample 1



Figure 11: Received Raw Data Sample 2

## **Data Pre-processing**

The data preprocessing steps were performed to better prepare the data for the predictive model and identifying risk factors. Since a large portion of the data set contained missing values, it was important to focus on correcting these instances. When examining the data, it was evident that some gas distributors kept extra records. Often there were variables that were recorded for a single company and all other companies had no records for those specific attributes. That means that some entries would exist for one attribute. In other words, 25% of the attribute would contain values while the other 75% were unknown. Having such sparse data would have complicated data analysis and, more importantly, risk analysis methods. In that event, the first data preprocessing step performed eliminated attributes that were deemed unnecessary for further computations. Eliminating attributes cleaned the data set up and assisted with the removal of a large number of missing values that existed. Furthermore, eliminating attributes actively reduces the dimensionality of the data set and allows for increased model performance with a reduced computation time. Data set reduction resulted in the following remaining attributes:

1. Incident Date
2. Underground Facility Operator
3. Latitude Coordinate
4. Longitude Coordinate

One of the goals of this study was to develop risk model to predict future risk involved on the gas pipe damage process which involve all parties. Thus, the distribution of 4 groups and the attributes was necessary to get precise results. It was concluded that additional attributes would be required in order to effectively reach the goal behind the research.

This concept was expanded to apply to nominal attributes which resulted in the following set of attributes:

1. Excavator Address
2. Damage Zip code
3. Locate Type Ticket
4. Latitude Coordinate
5. Locating Accuracy
6. Underground Facility Operator
7. Number of Excavating Request on the Same Area
8. Excavation Depth
9. Number of outgoing Calls
10. Number of Passed damages with 10-mile radius
11. Damage Description

In addition, the original nominal attributes were left in the raw data for easier identification, but were removed for any risk model application.

In addition, the data was distributed over 5 sheets in excel as 2010, 2011, 2012, 2013, and 2014. Then the data cleaned by some functions built in Excel. Then, some attribute was selected.

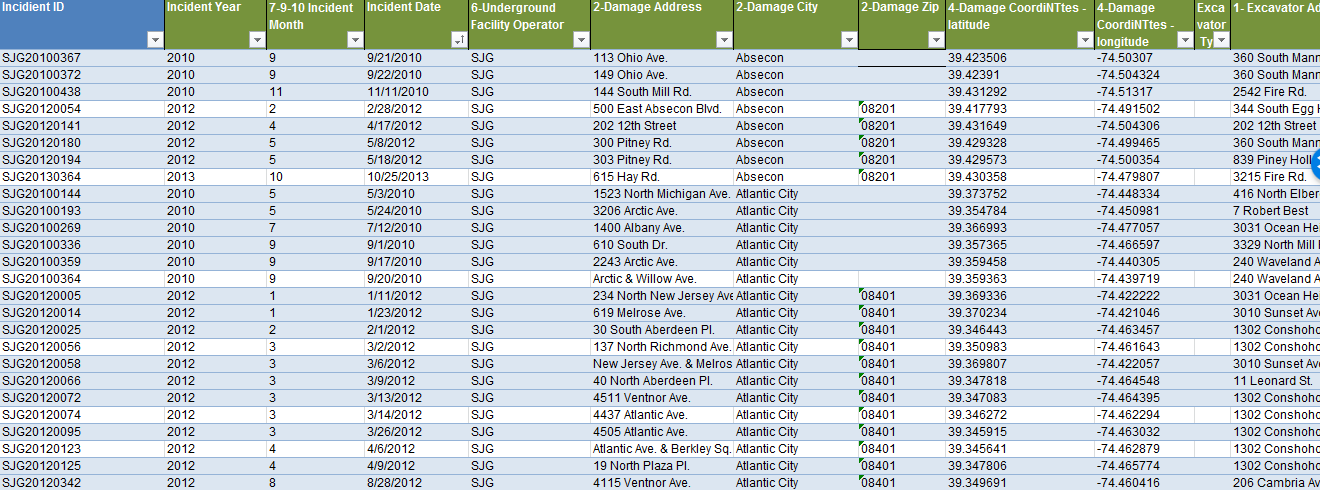


Figure 12: Organized data set sample

Then, the complete address was combined together including street number, name, city, state, and zip code. Then, excel functions were developed to cross reference the data set and produce two categories damaged data, and non-damaged data as in Figure 12.

## **Overview of Geo-code, Plotting, and Preliminary Exploratory Analysis of undamaged UG data**

### Step 1- Geo-code data, Latitude, and Longitude.

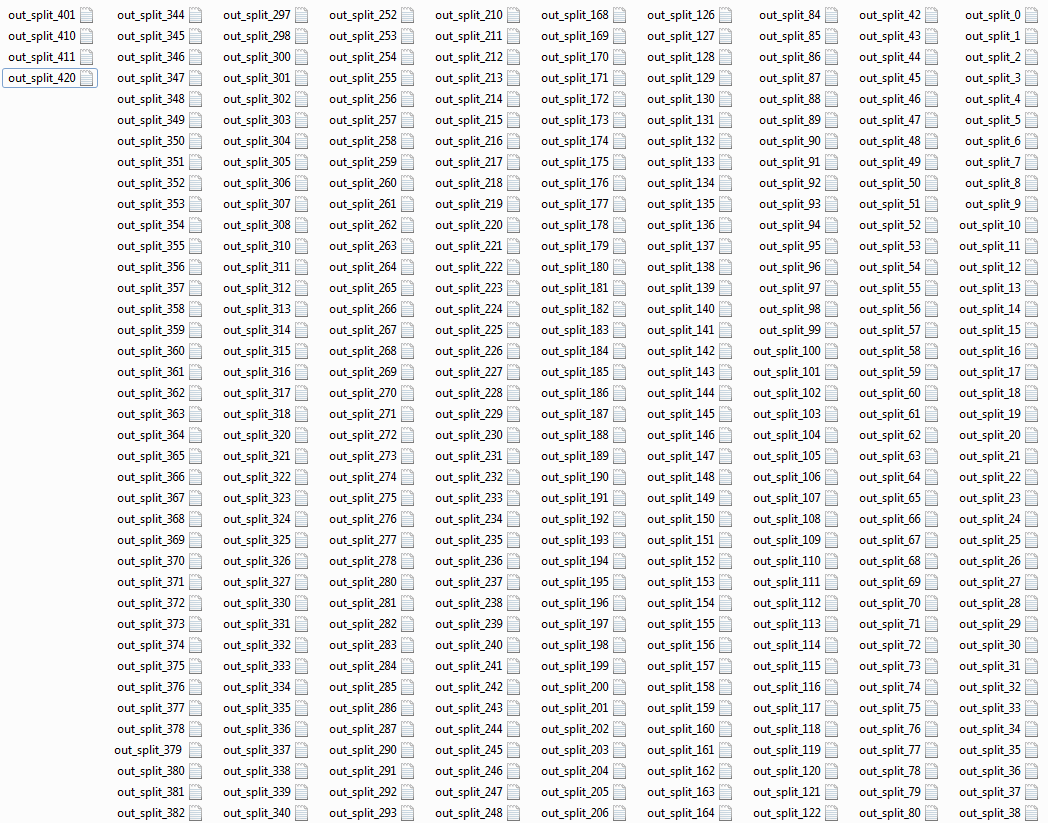
The collected undamaged data did not have the Latitude, and Longitude. Thus, it was not possible to plot large pile of data without Latitude, and Longitude. After determined the Latitude, and Longitude to all locations, the data was split into 420 files figure 00, each file contains complete address, and Lat/Long figure 14. Then, all files combined into one Excel file so we can import it to QGIS, OR ArcGIS later.

Figure 13 : Shows 420 files of undamaged data

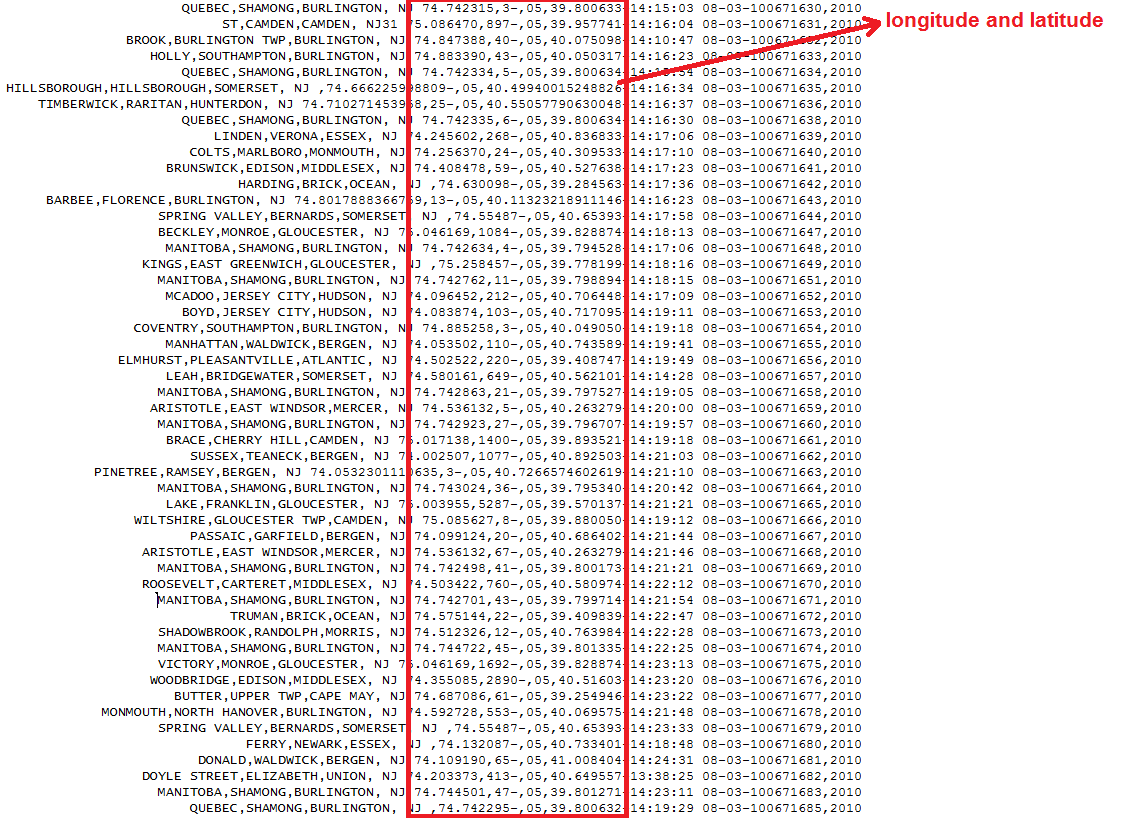


Figure 14 file shows latitude/Longitude of UG undamaged data

### Step 2- Create the map, and Plot all locations on the Map.

There is mutable software, platforms to perform the task of plotting locations. However, not all of them will handle large number of data, and requirement of spatial analysis including heat map. Thus, for the purpose of this research will use ArcGIS to plot, and analyze 775,000 locations of undamaged UG gas pipe data. In addition to, Fusion table was used to perform preliminary plotting as well.

All geocoded data was saved and combined to one CSV file. Then the CSV imported to ArcGIS through multiple steps including creating new project, creating new map, setting up layers, importing data to the map, and plotting .See figures 15,16.

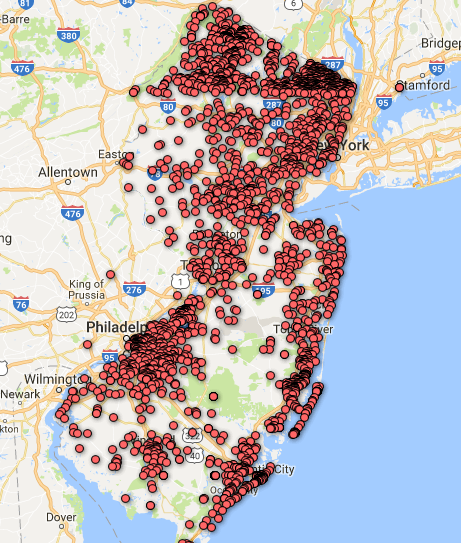


Figure 15 Shows plotted 775,000 locations using Fusion Table.

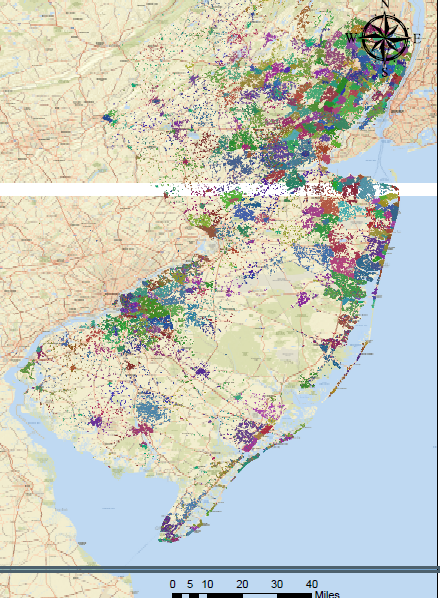


Figure 16 Shows plotted 775,000 locations using ArcGIS

## **Preparing Data Features/ Attributes**

The Data was all assembled in one file, as can be seen in figure 17. Each raw has a ticket number, date of the incident; either this incident was regular maintenance, or damage. The third column which is the time of incident was selected as important attribute, and it was available for damage, and undamaged data. The place of the ticket was identified and added to the data, which is city, and county. Cross referencing the damage for all data, every ticket request was checked if it was damaged or undamaged, for damaged (YES) was used, and for undamaged (NO) was used. In more details, (YES) means the one call center was called, and service was requested. However, the underground gas pipe was damaged. (NO) means the agency was called, and service was requested, and the maintenance was performed, UG gas pipe was not damaged.

In addition, the damaged tickets were received in different files than the undamaged data, thus, cross referencing, and compiling the data in one file was important task to perform before starting the data preparation.

As mentioned in the previous section, the damaged data was received with latitude, and longitude. However, undamaged data did not have latitude, and longitude. Thus, Server was built to determine the Latitude, and Longitude for the undamaged data, which is important feature in deriving new attributes, and specify the exact location for the underground gas pipeline. See figure 17 for columns ticket #, data, time, county, city, damage ( YES/NO), Longitude, and Latitude.

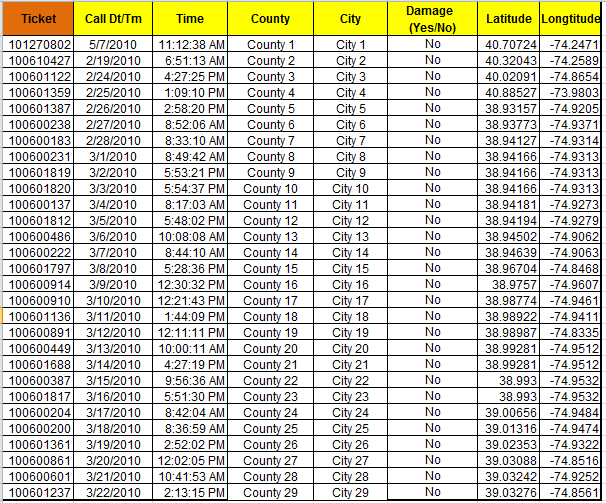


Figure 17: Assembeled Data ( Damage& undamage)

## **The following attributes were developed /derived**

1. Time (AM, or PM), the time of incident was categorized into two categories. Any time before 12 AM was categorized as equal to 1. Any time after 12 PM was categorized as PM equal to 1. Excel F function was used to map out the time into two categories AM, PM. See figure 18.

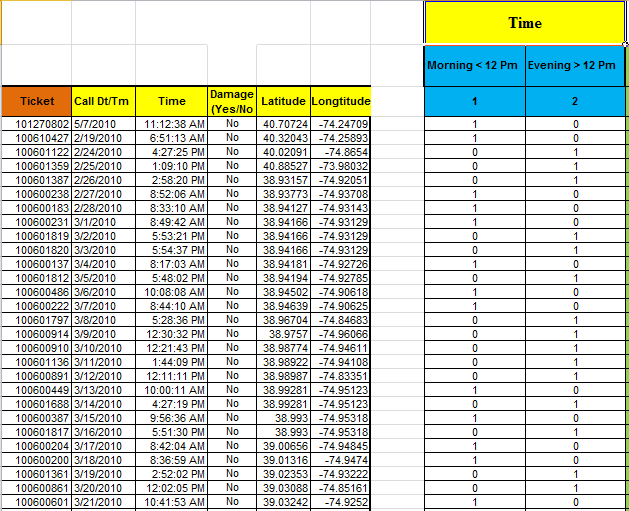


Figure 18 Time(AM/PM) Attribute

1. Week ( Mon, Tue, Wed, Thur, Fri, Sat, Sun), the week was selected as attribute, and categorized into 7 categories, so the function will look where is the ticket falling in and place the ticket in the day of the week which happened in. See figure 19.



Figure 19 Week ( Mon, Tue, Wed, Thu,Fri,Sat,Sun) Attribute

1. Month ( Jan, Feb, Mar, April, May, Jun, July, Aug, Sept, Oct, Nov, Dec), the Month was selected as attribute, and categorized into 12 categories, so the function will look where is the ticket falling in and place the ticket in the month of the year which happened in. See figure 20.



Figure 20: Months ( Jan, Feb,Mar, Apri, May, Jun, Jul, Aug, Sept, Oct, Nov, Dec)

1. Year (2010, 2011, 2012), the Year was selected as attribute, and categorized into 3 categories, so the function will look where is the ticket falling in and place the ticket in the Year which happened in. See figure 21.



Figure 21 YEARS ( 2010,2011, 2012)

1. Season (Spring, Summer, Fall, Winter), the seasons was selected as attribute, and categorized into 4 categories, so the function will look where is the ticket falling in and place the ticket in the right season which happened in. See figure 22.



Figure 22 Season ( Spring, Summer, Fall, Winter)

1. Location (County, City, Latitude, Longitude), the Location was selected as attribute, and categorized into 4 categories, so the function will look where is the ticket falling in and place the ticket in the right location within the right Lat, & Long which happened in. See figure 23.



Figure 23 : Location (County, City, Lat, Long)

1. Damages within 10 miles’ diameter, 20 miles’ diameter, 30 miles’ diameter (Diameter 10, Diameter 20, Diameter 30), the number of damages within different diameters for the same incident was selected as attribute, and categorized into 3 categories, Joint Spatial method used to calculate these values by using ARC GIS was us. See figure 24.

Joint spatial was introduce to transfer attributes from one layer to another based on their spatial relationship. Joint Spatial is a process which used in this study to transfer data from one feature layer's attribute to combined it with another layer's attributes. In addition, by developing this attribute, we should be able to see how surrounding damages within different density, and concentration with in certain distance could influence the damage of future underground gas pipeline.

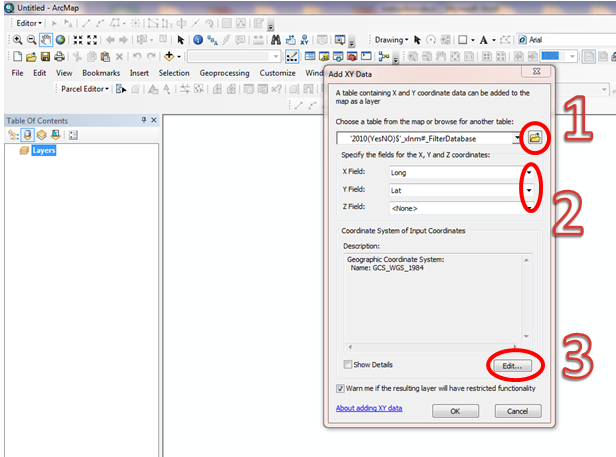
Steps 1. First step is to create XY events from excel file, then Browse for your excel file, set latitude, and longitude fields and set coordinate system to WGS 84, then OK. See figure 24.

Figure 24 Creating X,Y fileds in ARC GIS

Step 2. Plot the data on the map within provided Lat, and Long. There are some data with bad coordinate, see figure 25. This process is done as a part of a data cleaning, which is very important to get accurate and precise results. The locations which have bad coordinates needed to be removed from the data.

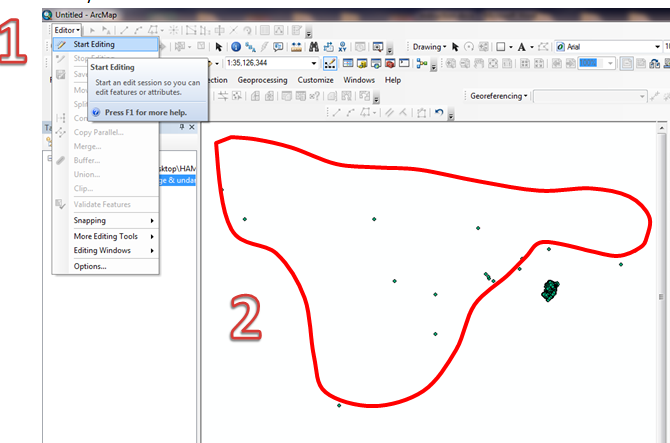


Figure 25Locating Bad Coordinates within the plotted Locations

Step 3. Create buffer areas of specific radius ( 10miles, 20 miles, 30 miles) Go to Search tab (CTRL+F) and type “Buffer” Select Buffer (Analysis) Tool and then we set the required parameters (Input Features, Output Features, Linear units set to miles and type10). Do the same thing for 20 and 30 miles separately. See figure 26.

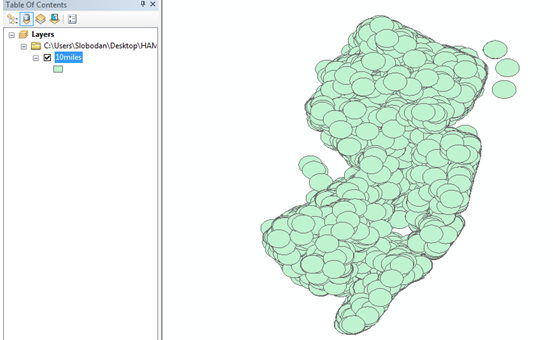


Figure 26: Buffer Circle 10 Miles Diameter

Step 4. Is this step we set the parameters, and select attributes, such as Damaged Yes, City, ticket, County, and merge between two layers, a new layer that contain Join\_Count field showing the number of locations inside 10 miles diameter area was created and joint spatial was conducted. See figure 27, 28.

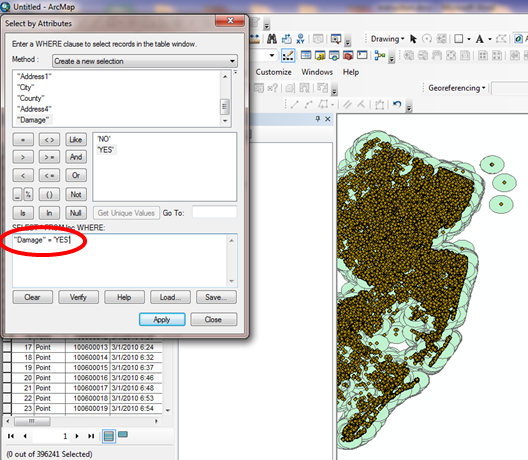


Figure 27 Joint Spatial Buffers within 10 miles Diameter

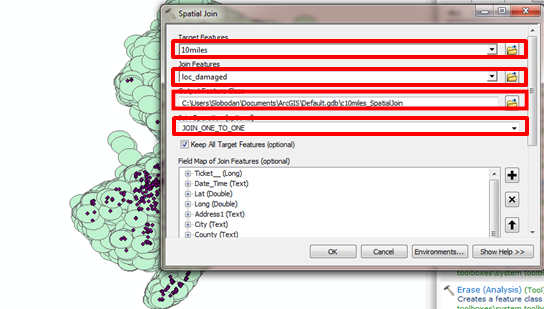


Figure 28 forming Joint Spatial

Step 5. Is this step we double check, and cross reference some of the damage numbers within certain ticket number. Finally, The results looks like ellipses because of projection, but actually these are circles See Figure 29. The results are ready to be exported to excel.. All five steps are repeated to determine the damages within 20 miles, 30 miles. See figure 30 for final outcome.

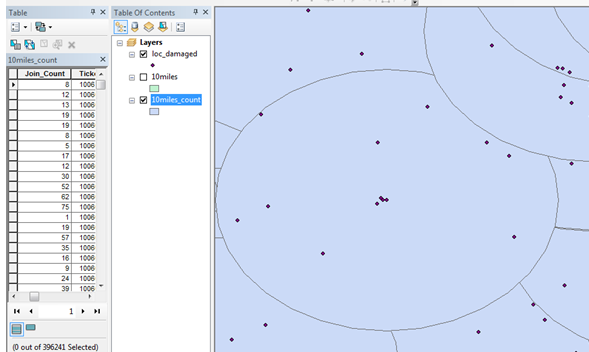


Figure 29 Magnified10 miles Circle.



Figure 30 Shows all damages within 10,20, 30-mile Diameter

## Univariate (Descriptive), and Bivariate Analysis

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words data has only one variable which is the time. It doesn’t deal with UG gas pipe damages or relationships (unlike regression) and its major purpose is to describe; it takes data, summarizes that data and finds patterns in the data. Some ways will be used to describe patterns found in Univariate attribute include central tendency (mean, mode and median) and dispersion: range, variance, maximum, minimum, quartiles (including the interquartile range), and standard deviation. Several options of describing data with Univariate outcome. Some techniques, Frequency Distribution Tables, Bar Charts, Histograms, Frequency Polygons, Pie Charts, Histograms. Bivariate analysis, analysis with two UG gas pipe attributes that can change and are compared to find relationships. If one variable is influencing another variable, then we have bivariate data that has an independent (UG gas pipe data attributes) and a dependent variable (Target, Yes, NO). This is because one variable depends on the other for change. An independent variable is a condition or piece of data in an experiment that can be controlled or changed. A dependent variables (Predictor) are a condition or piece of UG pipe data in an experiment that is controlled or influenced by an outside factor, most often the independent variable (which is in our case damage).

### Univariate ( Descriptive) Analysis ( Time AM/PM)

The time was converted from (AM/PM) before, and afternoon to 24 hrs (24hrs) (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24).

Table 2: Shows Descriptive Analysis for 24 hrs Attributes

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | | |
|  | **N** | **Range** | **Minimum** | **Maximum** | **Sum** | **Mean** | |
| **Statistic** | **Statistic** | **Statistic** | **Statistic** | **Statistic** | **Statistic** | **Std. Error** |
| **Time 24** | 396547 | 23 | 0 | 23 | 4786156 | 12.07 | .005 |
| **Valid N (lis twise)** | 396547 |  |  |  |  |  |  |

Table 3 Shows Discriptive Analysis ( St. Deviation, Skewness, Kurtosis)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | |
|  | **Std. Deviation** | **Variance** | **Skewness** | | **Kurtosis** | |
| **Statistic** | **Statistic** | **Statistic** | **Std. Error** | **Statistic** | **Std. Error** |
| **Time 24** | 3.319 | 11.016 | .270 | .004 | .085 | .008 |

Table 4 : Processing Summary 24 hrs

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Processing Summary 24 hrs** | | | | | | | |
|  | **Time 24** | **Cases** | | | | | |
| **Valid** | | **Missing** | | **Total** | |
| **N** | **Percent** | **N** | **Percent** | **N** | **Percent** |
| **DAMG** | **0** | 423 | 100.0% | 0 | 0.0% | 423 | 100.0% |
| **1** | 298 | 100.0% | 0 | 0.0% | 298 | 100.0% |
| **2** | 257 | 100.0% | 0 | 0.0% | 257 | 100.0% |
| **3** | 207 | 100.0% | 0 | 0.0% | 207 | 100.0% |
| **4** | 244 | 100.0% | 0 | 0.0% | 244 | 100.0% |
| **5** | 817 | 100.0% | 0 | 0.0% | 817 | 100.0% |
| **6** | 5945 | 100.0% | 0 | 0.0% | 5945 | 100.0% |
| **7** | 19740 | 100.0% | 0 | 0.0% | 19740 | 100.0% |
| **8** | 30985 | 100.0% | 0 | 0.0% | 30985 | 100.0% |
| **9** | 38866 | 100.0% | 0 | 0.0% | 38866 | 100.0% |
| **10** | 42844 | 100.0% | 0 | 0.0% | 42844 | 100.0% |
| **11** | 41272 | 100.0% | 0 | 0.0% | 41272 | 100.0% |
| **12** | 35185 | 100.0% | 0 | 0.0% | 35185 | 100.0% |
| **13** | 40578 | 100.0% | 0 | 0.0% | 40578 | 100.0% |
| **14** | 43808 | 100.0% | 0 | 0.0% | 43808 | 100.0% |
| **15** | 38610 | 100.0% | 0 | 0.0% | 38610 | 100.0% |
| **16** | 26111 | 100.0% | 0 | 0.0% | 26111 | 100.0% |
| **17** | 11467 | 100.0% | 0 | 0.0% | 11467 | 100.0% |
| **18** | 5762 | 100.0% | 0 | 0.0% | 5762 | 100.0% |
| **19** | 4323 | 100.0% | 0 | 0.0% | 4323 | 100.0% |
| **20** | 3426 | 100.0% | 0 | 0.0% | 3426 | 100.0% |
| **21** | 2753 | 100.0% | 0 | 0.0% | 2753 | 100.0% |
| **22** | 1789 | 100.0% | 0 | 0.0% | 1789 | 100.0% |
| **23** | 837 | 100.0% | 0 | 0.0% | 837 | 100.0% |

Table 5 Detailed Statistics analysis of Separate hrs from 8 hr-16 hr

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptive** | | | | | |
|  | Time 24 | | | Statistic | Std. Error |
|  | 8hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .01 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .005 |  |
| Std. Deviation | | .069 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 14.269 | .014 |
| Kurtosis | | 201.604 | .028 |
| 9hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .00 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .002 |  |
| Std. Deviation | | .050 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 20.047 | .012 |
| Kurtosis | | 399.908 | .025 |
| 10hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .00 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .003 |  |
| Std. Deviation | | .053 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 18.660 | .012 |
| Kurtosis | | 346.224 | .024 |
| 11hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .00 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .002 |  |
| Std. Deviation | | .050 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 19.943 | .012 |
| Kurtosis | | 395.750 | .024 |
| 12hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .00 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .001 |  |
| Std. Deviation | | .034 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 29.245 | .013 |
| Kurtosis | | 853.293 | .026 |
| 13hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .00 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .003 |  |
| Std. Deviation | | .054 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 18.308 | .012 |
| Kurtosis | | 333.194 | .024 |
| 14hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .00 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .003 |  |
| Std. Deviation | | .050 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 19.791 | .012 |
| Kurtosis | | 389.714 | .023 |
| 15hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .00 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .003 |  |
| Std. Deviation | | .055 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 17.930 | .012 |
| Kurtosis | | 319.498 | .025 |
| 16hr | Mean | | .00 | .000 |
| 95% Confidence Interval for Mean | Lower Bound | .00 |  |
| Upper Bound | .00 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .003 |  |
| Std. Deviation | | .052 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 19.100 | .015 |
| Kurtosis | | 362.833 | .030 |

### 

Figure 31 Shows Mean, Std. Deviation, And # of Frequencies

### Bivariate Analysis (Time AM/PM) & 24hr

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Time 24 \* DAMAGE Cross Tabulation** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
|  | 8hr | Count | 30835a | 150b | 30985 |
| Expected Count | 30902.7 | 82.3 | 30985.0 |
| % within Time 24 | 99.5% | 0.5% | 100.0% |
| % within DAMG | 7.8% | 14.2% | 7.8% |
| % of Total | 7.8% | 0.0% | 7.8% |
| 9hr | Count | 38770a | 96a | 38866 |
| Expected Count | 38762.8 | 103.2 | 38866.0 |
| % within Time 24 | 99.8% | 0.2% | 100.0% |
| % within DAMG | 9.8% | 9.1% | 9.8% |
| % of Total | 9.8% | 0.0% | 9.8% |
| 10hr | Count | 42722a | 122a | 42844 |
| Expected Count | 42730.2 | 113.8 | 42844.0 |
| % within Time 24 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 10.8% | 11.6% | 10.8% |
| % of Total | 10.8% | 0.0% | 10.8% |
| 11hr | Count | 41169a | 103a | 41272 |
| Expected Count | 41162.4 | 109.6 | 41272.0 |
| % within Time 24 | 99.8% | 0.2% | 100.0% |
| % within DAMG | 10.4% | 9.8% | 10.4% |
| % of Total | 10.4% | 0.0% | 10.4% |
| 12hr | Count | 35144a | 41b | 35185 |
| Expected Count | 35091.6 | 93.4 | 35185.0 |
| % within Time 24 | 99.9% | 0.1% | 100.0% |
| % within DAMG | 8.9% | 3.9% | 8.9% |
| % of Total | 8.9% | 0.0% | 8.9% |
| 13hr | Count | 40458a | 120a | 40578 |
| Expected Count | 40470.2 | 107.8 | 40578.0 |
| % within Time 24 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 10.2% | 11.4% | 10.2% |
| % of Total | 10.2% | 0.0% | 10.2% |
| 14hr | Count | 43697a | 111a | 43808 |
| Expected Count | 43691.7 | 116.3 | 43808.0 |
| % within Time 24 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 11.0% | 10.5% | 11.0% |
| % of Total | 11.0% | 0.0% | 11.0% |
| 15hr | Count | 38491a | 119a | 38610 |
| Expected Count | 38507.5 | 102.5 | 38610.0 |
| % within Time 24 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 9.7% | 11.3% | 9.7% |
| % of Total | 9.7% | 0.0% | 9.7% |
| 16hr | Count | 26040a | 71a | 26111 |
| Expected Count | 26041.7 | 69.3 | 26111.0 |
| % within Time 24 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 6.6% | 6.7% | 6.6% |
| % of Total | 6.6% | 0.0% | 6.6% |
| Total | | Count | 395494 | 1053 | 396547 |
| Expected Count | 395494.0 | 1053.0 | 396547.0 |
| % within Time 24 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |

|  |  |  |  |
| --- | --- | --- | --- |
| **P- Value (Chi-Square Tests**) | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 147.633a | 23 | .001 |
| Likelihood Ratio | 157.744 | 23 | .000 |
| Linear-by-Linear Association | .162 | 1 | .687 |
| McNemar - Bowker Test | . | . | .b |
| N of Valid Cases | 396547 |  |  |
| 1. 8 cells (16.7%) have expected count less than 5. The minimum expected count is .55. 2. Computed only for a PxP table, where P must be greater than 1. | | | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Directional Measures** | | | | | | | | | |
|  | | | | | Value | Asymptotic Standard Errora | | Approximate Tb | |
| Ordinal by Ordinal | Somers' d | | Symmetric | | .000 | .000 | | -.398 | |
| Time 24 Dependent | | -.007 | .018 | | -.398 | |
| DAMG Dependent | | .000 | .000 | | -.398 | |
|  | | | | | | | Approximate Significance | |
| Ordinal by Ordinal | | Somers' d | | Symmetric | | | .691 | |
| Time 24 Dependent | | | .691 | |
| DAMG Dependent | | | .691 | |
| a. Not assuming the null hypothesis. | | | | | | | | |
| b. Using the asymptotic standard error assuming the null hypothesis. | | | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Symmetric Measures** | | | | | | |
|  | | | Value | Asymptotic Standard Errora | | Approximate Tb |
| Ordinal by Ordinal | Kendall's tau-b | | -.001 | .001 | | -.398 |
| Kendall's tau-c | | .000 | .000 | | -.398 |
| Gamma | | -.008 | .020 | | -.398 |
| Spearman Correlation | | -.001 | .002 | | -.406 |
| Interval by Interval | Pearson's R | | -.001 | .002 | | -.402 |
| Measure of Agreement | Kappa | | .000 | .000 | | -.483 |
| N of Valid Cases | | | 396547 |  | |  |
| Approximate Significance | | | | | | | |
| Ordinal by Ordinal | | Kendall's tau-b | | | .691 | | |
| Kendall's tau-c | | | .691 | | |
| Gamma | | | .691 | | |
| Spearman Correlation | | | .684c | | |
| Interval by Interval | | Pearson's R | | | .687c | | |
| Measure of Agreement | | Kappa | | | .629 | | |
| N of Valid Cases | | | | |  | | |
| a. Not assuming the null hypothesis. | | | | | | | |
| b. Using the asymptotic standard error assuming the null hypothesis. | | | | | | | |
| c. Based on normal approximation. | | | | | | | |

### 

Figure 32 : Shows Frequency vs Damage

### Univaraite (Descriptive) Analysis Mon, Tue, Wed, Thu, Fri, Sat, and Sun (Week Days)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | |
|  | N | Minimum | Maximum | Sum | Mean | Std. Deviation |
| Statistic | Statistic | Statistic | Statistic | Statistic | Statistic |
| Mon | 396547 | 0 | 1 | 80355 | .20 | .402 |
| Tue | 396547 | 0 | 1 | 86437 | .22 | .413 |
| Wed | 396547 | 0 | 1 | 79225 | .20 | .400 |
| Thu | 396547 | 0 | 1 | 72204 | .18 | .386 |
| Fri | 396547 | 0 | 1 | 64705 | .16 | .370 |
| Sat | 396547 | 0 | 1 | 8824 | .02 | .148 |
| Sun | 396547 | 0 | 1 | 4797 | .01 | .109 |
| Valid N (list wise) | 396547 |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | |
|  | Variance | Skewness | | Kurtosis | |
| Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Mon | .162 | 1.480 | .004 | .189 | .008 |
| Tue | .170 | 1.366 | .004 | -.134 | .008 |
| Wed | .160 | 1.502 | .004 | .255 | .008 |
| Thu | .149 | 1.648 | .004 | .715 | .008 |
| Fri | .137 | 1.823 | .004 | 1.324 | .008 |
| Sat | .022 | 6.478 | .004 | 39.963 | .008 |
| Sun | .012 | 8.926 | .004 | 77.679 | .008 |
| Valid N (list wise) |  |  |  |  |  |

### Bivariate Analysis Mon, Tue, Wed, Thu, Fri, Sat, and Sun (Week Days)

The Bivariate analysis was conducted for all days. In order to find out which day is more significant in causing/contributing to underground gas pipe damage, P-Value was calculated for each day Separate starting from Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday.

### **Bivariate Analysis for Monday.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Bivariate Analysis Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Mon | 0 | Count | 315337a | 855a | 316192 |
| % within Mon | 99.7% | 0.3% | 100.0% |
| % within DAMG | 79.7% | 81.2% | 79.7% |
| % of Total | 79.5% | 0.2% | 79.7% |
| Standardized Residual | .0 | .5 |  |
| 1 | Count | 80157a | 198a | 80355 |
| % within Mon | 99.8% | 0.2% | 100.0% |
| % within DAMG | 20.3% | 18.8% | 20.3% |
| % of Total | 20.2% | 0.0% | 20.3% |
| Standardized Residual | .1 | -1.1 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Mon | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **P- Value (Chi-Square Tests**) | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 1.393a | 1 | .5942 |  |  |
| Continuity Correction | 1.304 | 1 | .253 |  |  |
| Likelihood Ratio | 1.419 | 1 | .234 |  |  |
| Fisher's Exact Test |  |  |  | .250 | .127 |
| Linear-by-Linear Association | 1.393 | 1 | .238 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 213.38. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |
| --- | --- |
| Odds ratio | 1.1887 |
| 95 % CI: | 0.6216 to 2.2860 |
| z statistic | 0.533 |
| Significance level | P = 0.5942 |

### 

Figure 33 Show total Ticket Numer/Yera, Total undamages in Monday, & Total Damages in Monday

### Bivariate Analysis for Tuesday.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Bivariate Analysis Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Tue | 0 | Count | 309289a | 821a | 310110 |
| % within Tue | 99.7% | 0.3% | 100.0% |
| % within DAMG | 78.2% | 78.0% | 78.2% |
| % of Total | 78.0% | 0.2% | 78.2% |
| Standardized Residual | .0 | -.1 |  |
| 1 | Count | 86205a | 232a | 86437 |
| % within Tue | 99.7% | 0.3% | 100.0% |
| % within DAMG | 21.8% | 22.0% | 21.8% |
| % of Total | 21.7% | 0.1% | 21.8% |
| Standardized Residual | .0 | .2 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Tue | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **P-Value (Chi-Square Tests**) | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .034a | 1 | .4251 |  |  |
| Continuity Correctionb | .022 | 1 | .883 |  |  |
| Likelihood Ratio | .034 | 1 | .854 |  |  |
| Fisher's Exact Test |  |  |  | .854 | .441 |
| Linear-by-Linear Association | .034 | 1 | .853 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 229.53. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |
| --- | --- |
| Odds ratio | 1.2941 |
| 95 % CI: | 0.6704 to 2.4387 |
| z statistic | 0.798 |
| Significance level | P = 0.4251 |

### 

Figure 34: Show total Ticket Number /Year, Total Ticket Number in Tuesday, & Total Damages in Tuesday

### 

### Bivariate Analysis for Wednesday.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Bivariate Analysis Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Wed | 0 | Count | 316490a | 832a | 317322 |
| % within Wed | 99.7% | 0.3% | 100.0% |
| % within DAMG | 80.0% | 79.0% | 80.0% |
| % of Total | 79.8% | 0.2% | 80.0% |
| Standardized Residual | .0 | -.4 |  |
| 1 | Count | 79004a | 221a | 79225 |
| % within Wed | 99.7% | 0.3% | 100.0% |
| % within DAMG | 20.0% | 21.0% | 20.0% |
| % of Total | 19.9% | 0.1% | 20.0% |
| Standardized Residual | .0 | .7 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Wed | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **P-Value (Chi-Square Tests)** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .672a | 1 | .3592 |  |  |
| Continuity Correctionb | .610 | 1 | .435 |  |  |
| Likelihood Ratio | .664 | 1 | .415 |  |  |
| Fisher's Exact Test |  |  |  | .419 | .217 |
| Linear-by-Linear Association | .672 | 1 | .412 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 210.38.  b. Computed only for a 2x2 table | | | | | |

|  |  |
| --- | --- |
| Odds ratio | 1.3454 |
| 95 % CI: | 0.7922 to 2.5125 |
| z statistic | 0.917 |
| Significance level | P = 0.3592 |

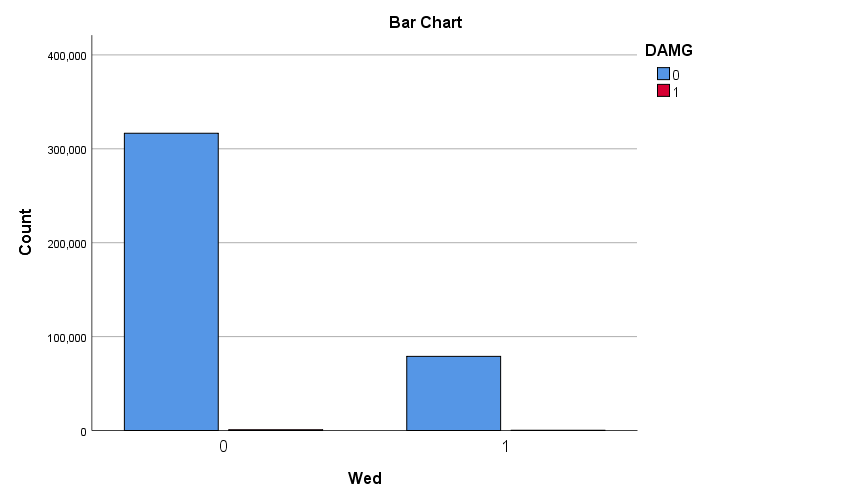


Figure 35: Show total Ticket Number /Year, Total Ticket Number in Wedensday, & Total Damages in Wedensday

### Bivariate Analysis for Thursday.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Thu | 0 | Count | 323486a | 857a | 324343 |
| % within Thu | 99.7% | 0.3% | 100.0% |
| % within DAMG | 81.8% | 81.4% | 81.8% |
| % of Total | 81.6% | 0.2% | 81.8% |
| Standardized Residual | .0 | -.1 |  |
| 1 | Count | 72008a | 196a | 72204 |
| % within Thu | 99.7% | 0.3% | 100.0% |
| % within DAMG | 18.2% | 18.6% | 18.2% |
| % of Total | 18.2% | 0.0% | 18.2% |
| Standardized Residual | .0 | .3 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Thu | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .116a | 1 | .4102 |  |  |
| Continuity Correctionb | .091 | 1 | .763 |  |  |
| Likelihood Ratio | .116 | 1 | .734 |  |  |
| Fisher's Exact Test |  |  |  | .722 | .382 |
| Linear-by-Linear Association | .116 | 1 | .733 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 191.73.  b. Computed only for a 2x2 table | | | | | |

|  |  |
| --- | --- |
| Odds ratio | 1.3063 |
| 95 % CI: | 0.6746 to 2.4869 |
| z statistic | 0.823 |
| Significance level | P = 0.4102 |

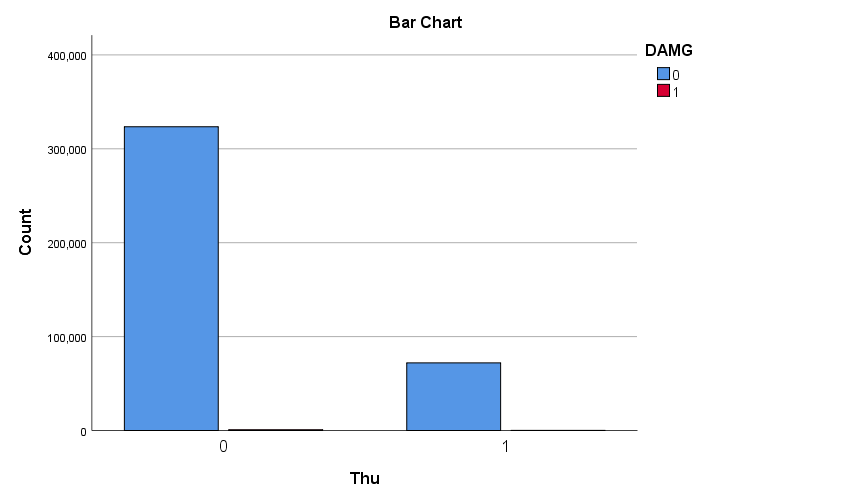


Figure 36: Show total Ticket Number /Year, Total Ticket Number in Thursday, & Total Damages in Thursday

* Bivariate Analysis for Friday

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Fri | 0 | Count | 330950a | 892a | 331842 |
| % within Fri | 99.7% | 0.3% | 100.0% |
| % within DAMG | 83.7% | 84.7% | 83.7% |
| % of Total | 83.5% | 0.2% | 83.7% |
| Standardized Residual | .0 | .4 |  |
| 1 | Count | 64544a | 161a | 64705 |
| % within Fri | 99.8% | 0.2% | 100.0% |
| % within DAMG | 16.3% | 15.3% | 16.3% |
| % of Total | 16.3% | 0.0% | 16.3% |
| Standardized Residual | .0 | -.8 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Fri | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .816a | 1 | .573 |  |  |
| Continuity Correctionb | .743 | 1 | .389 |  |  |
| Likelihood Ratio | .830 | 1 | .362 |  |  |
| Fisher's Exact Test |  |  |  | .380 | .194 |
| Linear-by-Linear Association | .816 | 1 | .366 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 171.82.  b. Computed only for a 2x2 table | | | | | |

|  |  |
| --- | --- |
| Odds ratio | 1.2250 |
| 95 % CI: | 0.6237 to 2.2037 |
| z statistic | 0.563 |
| Significance level | P = 0.5732 |

### 

Figure 37: Show total Ticket Number /Year, Total Ticket Number in Friday, & Total Damages in Friday

### Univerait (Descriptive) Analysis ( Months)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | | |
|  | N | Range | Minimum | Maximum | Sum | Mean | |
| Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error |
| Jan | 396547 | 0 | 0 | 0 | 0 | .00 | .000 |
| Feb | 396547 | 1 | 0 | 1 | 5 | .00 | .000 |
| Mar | 396547 | 1 | 0 | 1 | 45940 | .12 | .001 |
| Apr | 396547 | 1 | 0 | 1 | 53826 | .14 | .001 |
| May | 396547 | 1 | 0 | 1 | 47986 | .12 | .001 |
| Jun | 396547 | 1 | 0 | 1 | 50065 | .13 | .001 |
| Jul | 396547 | 1 | 0 | 1 | 42606 | .11 | .000 |
| Aug | 396547 | 1 | 0 | 1 | 44523 | .11 | .001 |
| Sep | 396547 | 1 | 0 | 1 | 35552 | .09 | .000 |
| Oct | 396547 | 1 | 0 | 1 | 23477 | .06 | .000 |
| Nov | 396547 | 1 | 0 | 1 | 33455 | .08 | .000 |
| Dec | 396547 | 1 | 0 | 1 | 19112 | .05 | .000 |
| Valid N (list wise) | 396547 |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | |
|  | Std. Deviation | Variance | Skewness | | Kurtosis | |
| Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Jan | .000 | .000 | . | . | . | . |
| Feb | .004 | .000 | 281.615 | .004 | 79305.400 | .008 |
| Mar | .320 | .102 | 2.401 | .004 | 3.763 | .008 |
| Apr | .343 | .117 | 2.127 | .004 | 2.524 | .008 |
| May | .326 | .106 | 2.324 | .004 | 3.402 | .008 |
| Jun | .332 | .110 | 2.251 | .004 | 3.065 | .008 |
| Jul | .310 | .096 | 2.535 | .004 | 4.428 | .008 |
| Aug | .316 | .100 | 2.456 | .004 | 4.033 | .008 |
| Sep | .286 | .082 | 2.873 | .004 | 6.253 | .008 |
| Oct | .236 | .056 | 3.735 | .004 | 11.954 | .008 |
| Nov | .278 | .077 | 2.991 | .004 | 6.945 | .008 |
| Dec | .214 | .046 | 4.219 | .004 | 15.799 | .008 |

### Bivariate Analysis ( Months)

The Bivariate analysis was conducted separate for all months. In order to find out which month is more significant in causing/contributing to underground gas pipe damage, P-Value was calculated for each month separate, starting from Jan, Feb, Mar, April, May, Jun, July, Aug, Sept, Oct, Nov, and Dec.

### Bivariate Analysis for March.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Mar | 0 | Count | 349599a | 1008b | 350607 |
| % within Mar | 99.7% | 0.3% | 100.0% |
| % within DAMG | 88.4% | 95.7% | 88.4% |
| % of Total | 88.2% | 0.3% | 88.4% |
| Standardized Residual | -.1 | 2.5 |  |
| 1 | Count | 45895a | 45b | 45940 |
| % within Mar | 99.9% | 0.1% | 100.0% |
| % within DAMG | 11.6% | 4.3% | 11.6% |
| % of Total | 11.6% | 0.0% | 11.6% |
| Standardized Residual | .4 | -7.0 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Mar | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 55.103a | 1 | .000 |  |  |
| Continuity Correctionb | 54.389 | 1 | .000 |  |  |
| Likelihood Ratio | 70.570 | 1 | .000 |  |  |
| Fisher's Exact Test |  |  |  | .000 | .000 |
| Linear-by-Linear Association | 55.103 | 1 | .000 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 121.99.  b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Mar (0 / 1) | .340 | .252 | .458 |
| For cohort DAMG = 0 | .998 | .998 | .998 |
| For cohort DAMG = 1 | 2.935 | 2.178 | 3.956 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for April.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Apr | 0 | Count | 341795a | 926a | 342721 |
| % within Apr | 99.7% | 0.3% | 100.0% |
| % within DAMG | 86.4% | 87.9% | 86.4% |
| % of Total | 86.2% | 0.2% | 86.4% |
| Standardized Residual | .0 | .5 |  |
| 1 | Count | 53699a | 127a | 53826 |
| % within Apr | 99.8% | 0.2% | 100.0% |
| % within DAMG | 13.6% | 12.1% | 13.6% |
| % of Total | 13.5% | 0.0% | 13.6% |
| Standardized Residual | .1 | -1.3 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Apr | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 2.060a | 1 | .151 |  |  |
| Continuity Correctionb | 1.933 | 1 | .164 |  |  |
| Likelihood Ratio | 2.128 | 1 | .145 |  |  |
| Fisher's Exact Test |  |  |  | .164 | .082 |
| Linear-by-Linear Association | 2.060 | 1 | .151 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 142.93. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Apr (0 / 1) | .873 | .725 | 1.051 |
| For cohort DAMG = 0 | 1.000 | .999 | 1.000 |
| For cohort DAMG = 1 | 1.145 | .952 | 1.378 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for May.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| May | 0 | Count | 347628a | 933a | 348561 |
| % within May | 99.7% | 0.3% | 100.0% |
| % within DAMG | 87.9% | 88.6% | 87.9% |
| % of Total | 87.7% | 0.2% | 87.9% |
| Standardized Residual | .0 | .2 |  |
| 1 | Count | 47866a | 120a | 47986 |
| % within May | 99.7% | 0.3% | 100.0% |
| % within DAMG | 12.1% | 11.4% | 12.1% |
| % of Total | 12.1% | 0.0% | 12.1% |
| Standardized Residual | .0 | -.7 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within May | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .493a | 1 | .482 |  |  |
| Continuity Correctionb | .429 | 1 | .512 |  |  |
| Likelihood Ratio | .502 | 1 | .479 |  |  |
| Fisher's Exact Test |  |  |  | .506 | .256 |
| Linear-by-Linear Association | .493 | 1 | .482 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 127.42.  b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for May (0 / 1) | .934 | .772 | 1.130 |
| For cohort DAMG = 0 | 1.000 | .999 | 1.000 |
| For cohort DAMG = 1 | 1.070 | .885 | 1.294 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for June.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Jun | 0 | Count | 345540a | 942b | 346482 |
| % within Jun | 99.7% | 0.3% | 100.0% |
| % within DAMG | 87.4% | 89.5% | 87.4% |
| % of Total | 87.1% | 0.2% | 87.4% |
| Standardized Residual | .0 | .7 |  |
| 1 | Count | 49954a | 111b | 50065 |
| % within Jun | 99.8% | 0.2% | 100.0% |
| % within DAMG | 12.6% | 10.5% | 12.6% |
| % of Total | 12.6% | 0.0% | 12.6% |
| Standardized Residual | .1 | -1.9 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Jun | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 4.156a | 1 | .041 |  |  |
| Continuity Correctionb | 3.969 | 1 | .046 |  |  |
| Likelihood Ratio | 4.370 | 1 | .037 |  |  |
| Fisher's Exact Test |  |  |  | .043 | .023 |
| Linear-by-Linear Association | 4.156 | 1 | .041 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 132.94. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Jun (0 / 1) | .815 | .669 | .992 |
| For cohort DAMG = 0 | .999 | .999 | 1.000 |
| For cohort DAMG = 1 | 1.226 | 1.008 | 1.492 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for July.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Jul | 0 | Count | 353013a | 928a | 353941 |
| % within Jul | 99.7% | 0.3% | 100.0% |
| % within DAMG | 89.3% | 88.1% | 89.3% |
| % of Total | 89.0% | 0.2% | 89.3% |
| Standardized Residual | .0 | -.4 |  |
| 1 | Count | 42481a | 125a | 42606 |
| % within Jul | 99.7% | 0.3% | 100.0% |
| % within DAMG | 10.7% | 11.9% | 10.7% |
| % of Total | 10.7% | 0.0% | 10.7% |
| Standardized Residual | -.1 | 1.1 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Jul | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 1.397a | 1 | .237 |  |  |
| Continuity Correctionb | 1.282 | 1 | .258 |  |  |
| Likelihood Ratio | 1.357 | 1 | .244 |  |  |
| Fisher's Exact Test |  |  |  | .231 | .129 |
| Linear-by-Linear Association | 1.397 | 1 | .237 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 113.14. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Jul (0 / 1) | 1.119 | .928 | 1.350 |
| For cohort DAMG = 0 | 1.000 | 1.000 | 1.001 |
| For cohort DAMG = 1 | .894 | .742 | 1.077 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for Aug.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Aug | 0 | Count | 351115a | 909b | 352024 |
| % within Aug | 99.7% | 0.3% | 100.0% |
| % within DAMG | 88.8% | 86.3% | 88.8% |
| % of Total | 88.5% | 0.2% | 88.8% |
| Standardized Residual | .0 | -.8 |  |
| 1 | Count | 44379a | 144b | 44523 |
| % within Aug | 99.7% | 0.3% | 100.0% |
| % within DAMG | 11.2% | 13.7% | 11.2% |
| % of Total | 11.2% | 0.0% | 11.2% |
| Standardized Residual | -.1 | 2.4 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Aug | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 6.346a | 1 | .012 |  |  |
| Continuity Correctionb | 6.102 | 1 | .014 |  |  |
| Likelihood Ratio | 5.983 | 1 | .014 |  |  |
| Fisher's Exact Test |  |  |  | .013 | .007 |
| Linear-by-Linear Association | 6.346 | 1 | .012 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 118.23.  b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Aug (0 / 1) | 1.253 | 1.051 | 1.495 |
| For cohort DAMG = 0 | 1.001 | 1.000 | 1.001 |
| For cohort DAMG = 1 | .798 | .670 | .952 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for Sept.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| * **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Sep | 0 | Count | 360084a | 911b | 360995 |
| % within Sep | 99.7% | 0.3% | 100.0% |
| % within DAMG | 91.0% | 86.5% | 91.0% |
| % of Total | 90.8% | 0.2% | 91.0% |
| Standardized Residual | .1 | -1.5 |  |
| 1 | Count | 35410a | 142b | 35552 |
| % within Sep | 99.6% | 0.4% | 100.0% |
| % within DAMG | 9.0% | 13.5% | 9.0% |
| % of Total | 8.9% | 0.0% | 9.0% |
| Standardized Residual | -.3 | 4.9 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Sep | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 26.428a | 1 | .000 |  |  |
| Continuity Correctionb | 25.876 | 1 | .000 |  |  |
| Likelihood Ratio | 23.221 | 1 | .000 |  |  |
| Fisher's Exact Test |  |  |  | .000 | .000 |
| Linear-by-Linear Association | 26.428 | 1 | .000 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 94.41. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Sep (0 / 1) | 1.585 | 1.328 | 1.892 |
| For cohort DAMG = 0 | 1.001 | 1.001 | 1.002 |
| For cohort DAMG = 1 | .632 | .530 | .754 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for Oct.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Oct | 0 | Count | 372107a | 963b | 373070 |
| % within Oct | 99.7% | 0.3% | 100.0% |
| % within DAMG | 94.1% | 91.5% | 94.1% |
| % of Total | 93.8% | 0.2% | 94.1% |
| Standardized Residual | .0 | -.9 |  |
| 1 | Count | 23387a | 90b | 23477 |
| % within Oct | 99.6% | 0.4% | 100.0% |
| % within DAMG | 5.9% | 8.5% | 5.9% |
| % of Total | 5.9% | 0.0% | 5.9% |
| Standardized Residual | -.2 | 3.5 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Oct | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 13.078a | 1 | .000 |  |  |
| Continuity Correctionb | 12.610 | 1 | .000 |  |  |
| Likelihood Ratio | 11.590 | 1 | .001 |  |  |
| Fisher's Exact Test |  |  |  | .000 | .000 |
| Linear-by-Linear Association | 13.078 | 1 | .000 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 62.34.  b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Oct (0 / 1) | 1.487 | 1.198 | 1.846 |
| For cohort DAMG = 0 | 1.001 | 1.000 | 1.002 |
| For cohort DAMG = 1 | .673 | .543 | .835 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for Nov.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Nov | 0 | Count | 362132a | 960a | 363092 |
| % within Nov | 99.7% | 0.3% | 100.0% |
| % within DAMG | 91.6% | 91.2% | 91.6% |
| % of Total | 91.3% | 0.2% | 91.6% |
| Standardized Residual | .0 | -.1 |  |
| 1 | Count | 33362a | 93a | 33455 |
| % within Nov | 99.7% | 0.3% | 100.0% |
| % within DAMG | 8.4% | 8.8% | 8.4% |
| % of Total | 8.4% | 0.0% | 8.4% |
| Standardized Residual | .0 | .4 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Nov | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |
| --- |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **P-Value (Chi-Square Tests**) | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .214a | 1 | .644 |  |  |
| Continuity Correctionb | .165 | 1 | .684 |  |  |
| Likelihood Ratio | .211 | 1 | .646 |  |  |
| Fisher's Exact Test |  |  |  | .622 | .342 |
| Linear-by-Linear Association | .214 | 1 | .644 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 88.84. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Nov (0 / 1) | 1.052 | .850 | 1.301 |
| For cohort DAMG = 0 | 1.000 | 1.000 | 1.001 |
| For cohort DAMG = 1 | .951 | .769 | 1.176 |
| N of Valid Cases | 396547 |  |  |

### Bivariate Analysis for Dec.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Dec | 0 | Count | 376438a | 997a | 377435 |
| % within Dec | 99.7% | 0.3% | 100.0% |
| % within DAMG | 95.2% | 94.7% | 95.2% |
| % of Total | 94.9% | 0.3% | 95.2% |
| Standardized Residual | .0 | -.2 |  |
| 1 | Count | 19056a | 56a | 19112 |
| % within Dec | 99.7% | 0.3% | 100.0% |
| % within DAMG | 4.8% | 5.3% | 4.8% |
| % of Total | 4.8% | 0.0% | 4.8% |
| Standardized Residual | .0 | .7 |  |
| Total | | Count | 395494 | 1053 | 396547 |
| % within Dec | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .572a | 1 | .449 |  |  |
| Continuity Correctionb | .468 | 1 | .494 |  |  |
| Likelihood Ratio | .554 | 1 | .457 |  |  |
| Fisher's Exact Test |  |  |  | .432 | .247 |
| Linear-by-Linear Association | .572 | 1 | .449 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 50.75. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Dec (0 / 1) | 1.110 | .847 | 1.453 |
| For cohort DAMG = 0 | 1.000 | 1.000 | 1.001 |
| For cohort DAMG = 1 | .902 | .689 | 1.180 |
| N of Valid Cases | 396547 |  |  |

### Univariate ( Descriptive) Analysis (Seasons)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptive** | | | | | |
|  | Season | | | Statistic | Std. Error |
| Damage binary | Autumn |  | |  |  |
| Variance | | .004 |  |
| Std. Deviation | | .059 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 16.780 | .008 |
| Kurtosis | | 279.585 | .016 |
| Spring |  | |  |  |
| Variance | | .002 |  |
| Std. Deviation | | .044 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 22.428 | .006 |
| Kurtosis | | 501.019 | .013 |
| Summer |  | |  |  |
| Variance | | .003 |  |
| Std. Deviation | | .053 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 18.922 | .007 |
| Kurtosis | | 356.053 | .013 |
| Winter | Variance | | .003 |  |
| Std. Deviation  Minimum | .054 |  |  |
| 0 |  |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 18.396 | .018 |
| Kurtosis | | 336.466 | .035 |

### Bivariate Analysis Seasons ( Autumn, Spring, Summer, Winter)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Season \* DAMG Cross tabulation** | | | | |
| Count | | | | |
|  | | DAMG | | Total |
| 0 | 1 |
| Season | Autumn | 92159a | 325b | 92484 |
| Spring | 147460a | 292b | 147752 |
| Summer | 136814a | 380a | 137194 |
| Winter | 19061a | 56a | 19117 |
| Total | | 395494 | 1053 | 396547 |
| Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level. | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests for all Seasons Combined** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 52.701a | 3 | .000 |
| Likelihood Ratio | 52.801 | 3 | .000 |
| N of Valid Cases | 396547 |  |  |

### 

Figure 38 Total Undamaged & Damages per Season

### **Spring P Value**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 41.013a | 1 | .000 |  |  |
| Continuity Correctionb | 40.606 | 1 | .000 |  |  |
| Likelihood Ratio | 42.815 | 1 | .000 |  |  |
| Fisher's Exact Test |  |  |  | .000 | .000 |
| Linear-by-Linear Association | 41.013 | 1 | .000 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Spring (0 / 1) | .645 | .564 | .739 |
| For cohort DAMG = 0 | .999 | .999 | .999 |
| For cohort DAMG = 1 | 1.548 | 1.353 | 1.771 |
| N of Valid Cases | 396547 |  |  |

### **Summer P Value**

### 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 1.036a | 1 | .309 |  |  |
| Continuity Correctionb | .971 | 1 | .324 |  |  |
| Likelihood Ratio | 1.029 | 1 | .310 |  |  |
| Fisher's Exact Test |  |  |  | .315 | .162 |
| Linear-by-Linear Association | 1.036 | 1 | .309 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Summer (0 / 1) | 1.068 | .941 | 1.211 |
| For cohort DAMG = 0 | 1.000 | 1.000 | 1.001 |
| For cohort DAMG = 1 | .937 | .826 | 1.062 |
| N of Valid Cases | 396547 |  |  |

### **Autumn P Value**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 33.582a | 1 | .000 |  |  |
| Continuity Correctionb | 33.160 | 1 | .000 |  |  |
| Likelihood Ratio | 31.459 | 1 | .000 |  |  |
| Fisher's Exact Test |  |  |  | .000 | .000 |
| Linear-by-Linear Association | 33.581 | 1 | .000 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Autumn (0 / 1) | 1.469 | 1.289 | 1.675 |
| For cohort DAMG = 0 | 1.001 | 1.001 | 1.002 |
| For cohort DAMG = 1 | .681 | .598 | .776 |
| N of Valid Cases | 396547 |  |  |

### **Winter P Value**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .569a | 1 | .451 |  |  |
| Continuity Correctionb | .466 | 1 | .495 |  |  |
| Likelihood Ratio | .551 | 1 | .458 |  |  |
| Fisher's Exact Test |  |  |  | .432 | .248 |
| Linear-by-Linear Association | .569 | 1 | .451 |  |  |
| N of Valid Cases | 396547 |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for Winter (0 / 1) | 1.109 | .847 | 1.452 |
| For cohort DAMG = 0 | 1.000 | 1.000 | 1.001 |
| For cohort DAMG = 1 | .902 | .689 | 1.180 |
| N of Valid Cases | 396547 |  |  |

### Univariate ( Descriptive) Analysis ( County)

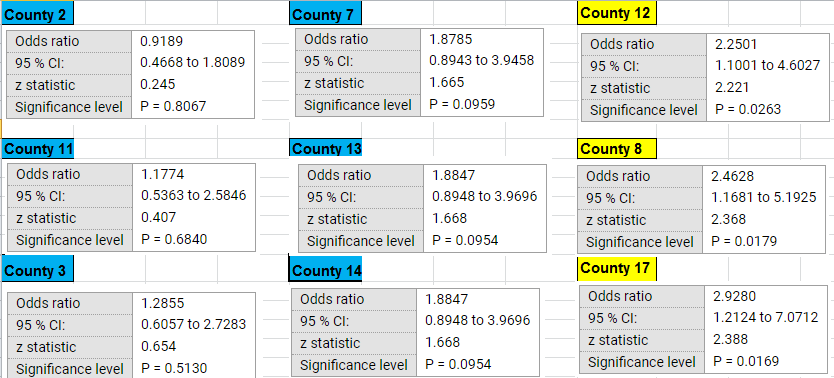


Figure 39:Shows counties with P Value & Odd Ratio

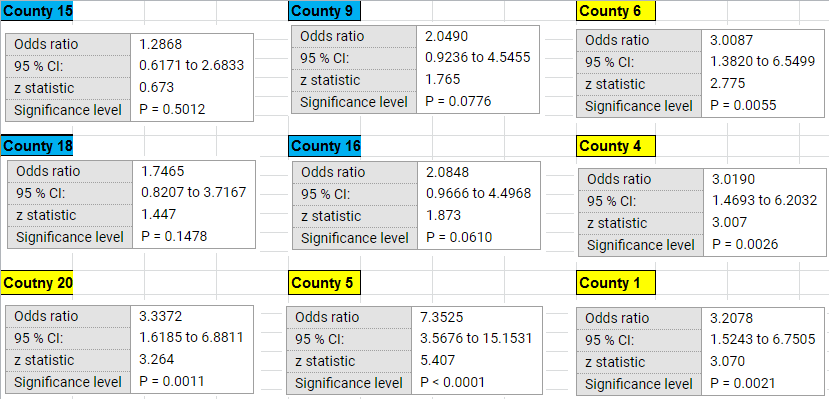


Figure 40: Shows sample of counties with P Value & Odd Ratio

Table 6: Shows The Risk Ratio, Odds Ratio, and P Value



### Univariate ( Descriptive) Analysis ( D10, D20, D30)

The input for this attribute was developed by geocoding the data, then plotting all data in ARC GIS, and after multiple processes including Joint Spatial, damages were calculated within 10 miles diameter. Processes were repeated for 20 miles, and 30 miles diameter. Because all values are continuous values, and it cannot be statistically analyzed, all values were converted into ranges, each range represent certain number of damages, then these ranges entered into SPSS for the statistics analysis. The tables below show only 50 samples of the total number see Table 5. The Ranges were classified by distributing the number of total damages into equal intervals of 4 ranges as following;

* Damages within 10 miles diameter
* Divided into Ranges ( Range 1 from 0-20 damages)
* Divided into Ranges ( Range 2 from 21-40 damages)
* Divided into Ranges ( Range 3 from 41-60 damages)
* Divided into Ranges ( Range 4 from 61-86damages)
* Damages within 20 miles diameter
* Divided into Ranges ( Range 1 from 0 - 49 damages)
* Divided into Ranges ( Range 2 from 50 - 98 damages)
* Divided into Ranges ( Range 3 from 99 - 147 damages)
* Divided into Ranges ( Range 4 from 148 – 196 damages)
* Damages within 30 miles diameter
* Divided into Ranges ( Range 1 from 0 - 73 damages)
* Divided into Ranges ( Range 2 from 74 - 147 damages)
* Divided into Ranges ( Range 3 from 148 - 222 damages)
* Divided into Ranges ( Range 4 from 223 – 295 damages)

Table 7: Classifying D10, D20, D30 into Four Ranges



|  |  |  |
| --- | --- | --- |
| **Statistics** | | |
| **D10** | | |
| N | Valid | 396546 |
| Missing | 1 |
| Mean | | 22.95 |
| Median | | 18.00 |
| Std. Deviation | | 18.422 |
| Range | | 86 |
| Minimum | | 0 |
| Maximum | | 86 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case Processing Summary** | | | | | | |
|  | Cases | | | | | |
| Included | | Excluded | | Total | |
| N | Percent | N | Percent | N | Percent |
| D10 | 396546 | 100.0% | 1 | 0.0% | 396547 | 100.0% |

### Bivariate Analysis ( D10).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DAMAGE within 10 mile Diameter Cross Tabulation** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Damage | 1 | Count | 225180 | 554 | 225734 |
| % with in low to high level of near area damage | 99.8% | 0.2% | 100.0% |
| % within DAMG | 56.9% | 52.6% | 56.9% |
| % of Total | 56.8% | 0.1% | 56.9% |
| 2 | Count | 116167 | 311 | 116478 |
| % within low to high level of near area damage | 99.7% | 0.3% | 100.0% |
| % within DAMG | 29.4% | 29.5% | 29.4% |
| % of Total | 29.3% | 0.1% | 29.4% |
| 3 | Count | 26929 | 89 | 27018 |
| % within low to high level of near area damage | 99.7% | 0.3% | 100.0% |
| % within DAMG | 6.8% | 8.5% | 6.8% |
| % of Total | 6.8% | 0.0% | 6.8% |
| 4 | Count | 27217 | 99 | 27316 |
| % within low to high level of near area damage | 99.6% | 0.4% | 100.0% |
| % within DAMG | 6.9% | 9.4% | 6.9% |
| % of Total | 6.9% | 0.0% | 6.9% |
| Total | | Count | 395493 | 1053 | 396546 |
| % within low to high level of near area damage | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |

|  |  |  |  |
| --- | --- | --- | --- |
| **P Value / Chi-Square Tests** **(D 10)** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 17.303a | 3 | .001 |
| Likelihood Ratio | 16.097 | 3 | .001 |
| Linear-by-Linear Association | 16.296 | 1 | .000 |
| N of Valid Cases | 396546 |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Symmetric Measures** | | | | |
|  | | Value | Asymptotic Standard Errora | Approximate Tb |
| Nominal by Nominal | Phi | .007 |  |  |
| Cramer's V | .007 |  |  |
| Interval by Interval | Pearson's R | .006 | .002 | 4.037 |
| Ordinal by Ordinal | Spearman Correlation | .006 | .002 | 3.527 |
| N of Valid Cases | | 396546 |  |  |

|  |  |  |
| --- | --- | --- |
| Approximate Significance | | |
| Nominal by Nominal | Phi | .001 |
| Cramer's V | .001 |
| Interval by Interval | Pearson's R | .000c |
| Ordinal by Ordinal | Spearman Correlation | .000c |
| N of Valid Cases | |  |
| a. Not assuming the null hypothesis. | | |
| b. Using the asymptotic standard error assuming the null hypothesis. | | |
| c. Based on normal approximation. | | |

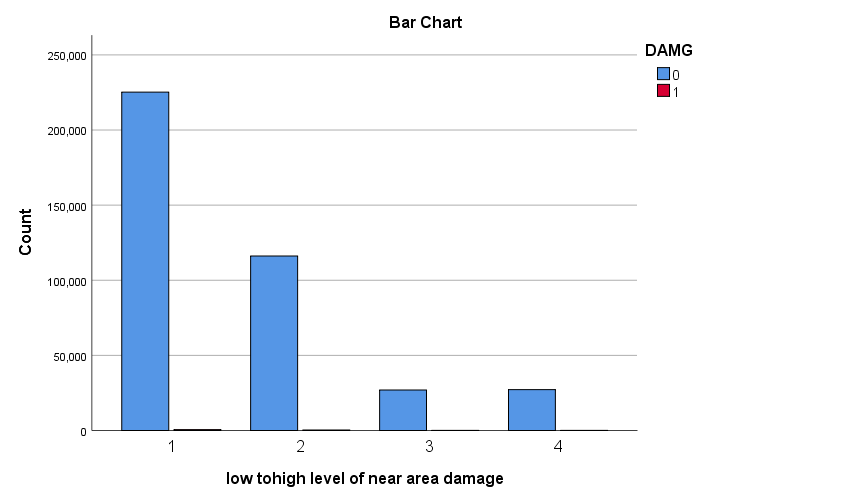


Figure 41 : Shows classification of Damages, and Undamaged per Range (D10)

### Univariate (Descriptive) Analysis (D20)

|  |  |  |
| --- | --- | --- |
| **Statistics** | | |
| **D20** | | |
| N | Valid | 396546 |
| Missing | 1 |
| Mean | | 73.93 |
| Median | | 62.00 |
| Std. Deviation | | 47.487 |
| Range | | 196 |
| Minimum | | 0 |
| Maximum | | 196 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case Processing Summary** | | | | | | |
|  | Cases | | | | | |
| Valid | | Missing | | Total | |
| N | Percent | N | Percent | N | Percent |
| low to high D20 \* DAMG | 396546 | 100.0% | 1 | 0.0% | 396547 | 100.0% |

### Bivariate Analysis (D20)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Low to High D20 \* DAMAGE Cross Tabulation** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| low to high D20 | 1.00 | Count | 154390 | 375 | 154765 |
| % within low to high D20 | 99.8% | 0.2% | 100.0% |
| % within DAMG | 39.0% | 35.6% | 39.0% |
| % of Total | 38.9% | 0.1% | 39.0% |
| 2.00 | Count | 119629 | 314 | 119943 |
| % within low to high D20 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 30.2% | 29.8% | 30.2% |
| % of Total | 30.2% | 0.1% | 30.2% |
| 3.00 | Count | 85125 | 256 | 85381 |
| % within low to high D20 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 21.5% | 24.3% | 21.5% |
| % of Total | 21.5% | 0.1% | 21.5% |
| 4.00 | Count | 36349 | 108 | 36457 |
| % within low to high D20 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 9.2% | 10.3% | 9.2% |
| % of Total | 9.2% | 0.0% | 9.2% |
| Total | | Count | 395493 | 1053 | 396546 |
| % within low to high D20 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square/ P Value Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 8.308a | 3 | .040 |
| Likelihood Ratio | 8.204 | 3 | .042 |
| Linear-by-Linear Association | 7.512 | 1 | .006 |
| N of Valid Cases | 396546 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 96.81. | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symmetric Measures** | | | | | |
|  | | | Value | Asymptotic Standard Errora | Approximate Tb |
| Nominal by Nominal | Phi | | .005 |  |  |
| Cramer's V | | .005 |  |  |
| Interval by Interval | Pearson's R | | .004 | .002 | 2.741 |
| Ordinal by Ordinal | Spearman Correlation | | .004 | .002 | 2.775 |
| N of Valid Cases | | | 396546 |  |  |
| Approximate Significance | | | | | |
| Nominal by Nominal | | Phi | | | .040 |
| Cramer's V | | | .040 |
| Interval by Interval | | Pearson's R | | | .006c |
| Ordinal by Ordinal | | Spearman Correlation | | | .006c |
| N of Valid Cases | | | | |  |
| a. Not assuming the null hypothesis. | | | | | |
| b. Using the asymptotic standard error assuming the null hypothesis. | | | | | |
| c. Based on normal approximation. | | | | | |

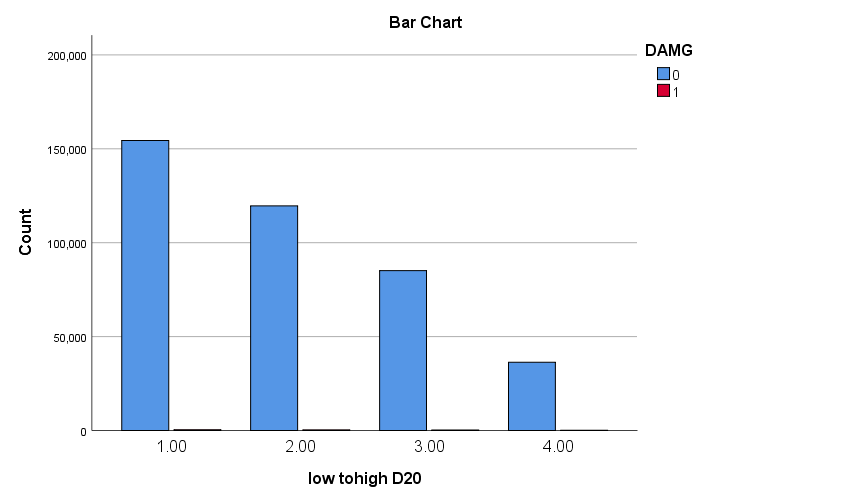


Figure 42Figure 24 : Show classification of Damages, and Undamaged per Range (D20)

### Univariate (Descriptive) Analysis (D30)

|  |  |  |
| --- | --- | --- |
| **D30** | | |
| N | Valid | 396546 |
| Missing | 1 |
| Mean | | 136.01 |
| Median | | 128.00 |
| Std. Deviation | | 75.655 |
| Range | | 295 |
| Minimum | | 0 |
| Maximum | | 295 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case Processing Summary** | | | | | | |
|  | Cases | | | | | |
| Valid | | Missing | | Total | |
| N | Percent | N | Percent | N | Percent |
| Low to High D 30 \* DAMG | 396546 | 100.0% | 1 | 0.0% | 396547 | 100.0% |

### Bivariate Analysis (D30)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Low to High D 30 \* DAMG Cross Tabulation** | | | | | |
|  | | | DAMG | | Total |
| 0 | 1 |
| Low to High D 30 | 1.00 | Count | 98560 | 266 | 98826 |
| % within Low to High D 30 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 24.9% | 25.3% | 24.9% |
| % of Total | 24.9% | 0.1% | 24.9% |
| 2.00 | Count | 126567 | 308 | 126875 |
| % within Low to High D 30 | 99.8% | 0.2% | 100.0% |
| % within DAMG | 32.0% | 29.2% | 32.0% |
| % of Total | 31.9% | 0.1% | 32.0% |
| 3.00 | Count | 99066 | 280 | 99346 |
| % within Low to High D 30 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 25.0% | 26.6% | 25.1% |
| % of Total | 25.0% | 0.1% | 25.1% |
| 4.00 | Count | 71300 | 199 | 71499 |
| % within Low to High D 30 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 18.0% | 18.9% | 18.0% |
| % of Total | 18.0% | 0.1% | 18.0% |
| Total | | Count | 395493 | 1053 | 396546 |
| % within Low to High D 30 | 99.7% | 0.3% | 100.0% |
| % within DAMG | 100.0% | 100.0% | 100.0% |
| % of Total | 99.7% | 0.3% | 100.0% |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square/ P Value Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 3.974a | 3 | .264 |
| Likelihood Ratio | 4.021 | 3 | .259 |
| Linear-by-Linear Association | .834 | 1 | .361 |
| N of Valid Cases | 396546 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 189.86. | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symmetric Measures** | | | | | |
|  | | | Value | Asymptotic Standard Errora | Approximate Tb |
| Nominal by Nominal | Phi | | .003 |  |  |
| Cramer's V | | .003 |  |  |
| Interval by Interval | Pearson's R | | .001 | .002 | .913 |
| Ordinal by Ordinal | Spearman Correlation | | .001 | .002 | .906 |
| N of Valid Cases | | | 396546 |  |  |
| Approximate Significance | | | | | |
| Nominal by Nominal | | Phi | | | .264 |
| Cramer's V | | | .264 |
| Interval by Interval | | Pearson's R | | | .361c |
| Ordinal by Ordinal | | Spearman Correlation | | | .365c |
| N of Valid Cases | | | | |  |
| a. Not assuming the null hypothesis. | | | | | |
| b. Using the asymptotic standard error assuming the null hypothesis. | | | | | |
| c. Based on normal approximation. | | | | | |

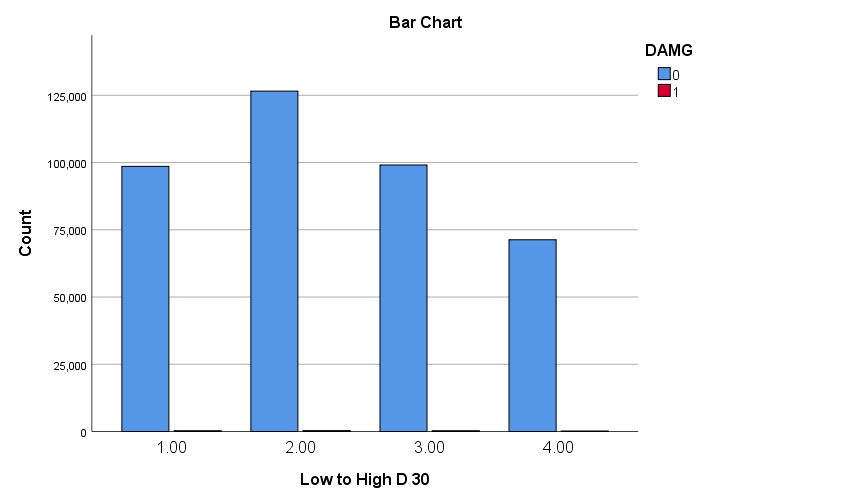


Figure 26 Show classification of Damages, and Undamaged per Range (D30)

## Discussion of Results, and Future Work.

## **Time,** first the time was divided into (AM, PM), and then divided to 24 hrs. Univariate and Bivariate analysis was performed. Findings, Timing attribute of the digging was determined to have significant effect on the UG gas pipe damage. Moreover, when all time analyzed as one attribute, P-Value was 0.001 which is less than the industry standard of 0.05. More specifically, 24hrs were analyzed separately; some hours have more damage percent than others. For, Instance, the damage percent within hour 8 represent 14.2% as compared to the rest of the 24 hrs, which is the highest percent of damage among all hours. In addition, the damage percent within hour 10, 13, and 15 represent 11.6%, 11.4%, 11.3% consecutively as compared to the rest of the 24 hrs. Therefore, the attribute of damage has significant impact on the UG gas pipe damage. In Summary, the received UG gas pipe digging requests by the agency during 10, 13, and 15 have more chance of been damaged than the other hours. Thus, by avoiding digging when it's possible during these hours, UG gas pipe damage risk will be decreased.

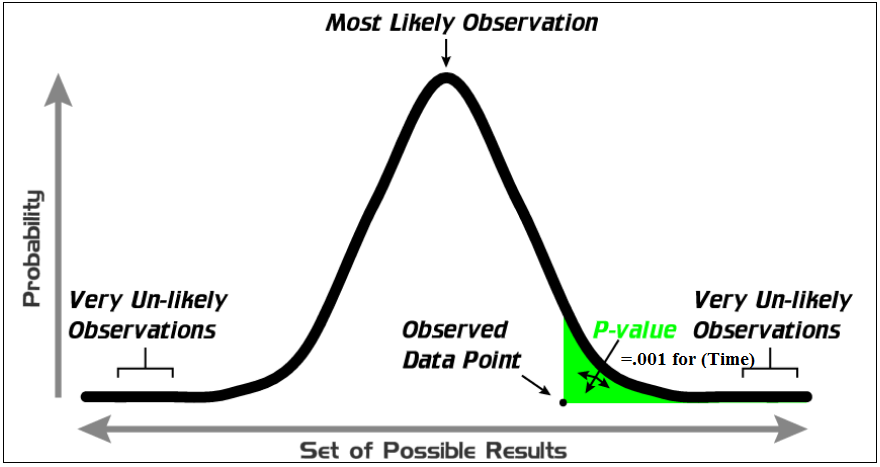


Figure 43: P Value for the attribute of (Time)

**Week,** First, the week as one attribute was determined to be significant. However, as observed form the UG gas pipe damaged data distribute, the damage was happening across all week days. Thus, Univaraite, and Bivariate statistics analysis was performed on each day separate to see the impact for each individual day on the UG gas pipe damage. Moreover, as per conducted analysis, one days was having significant impact on UG gas pipe damage, Saturday, , P values were 0.045. Thus, by avoiding digging operation on Saturday, that will significantly decrease the damage of UG gas pipe damage. As can be seen in the detailed analysis above, P value for the rest of the week days was more than 0.05 which means it does not have significant impact. In addition, odds ratio for Thursday, and Wednesday was 1.3. This means these two days have 1.3 chance of getting damaged as compared to rest of the week days. Results, any received digging requests during Thursday, and Wednesday will have 30% chance of been damage than the other days.

**Months,** The Bivariate analysis was conducted separate for all months together. In order to find out which month is more significant in causing/contributing to underground gas pipe damage, P-Value was calculated for each month separate, starting from Jan, Feb, Mar, April, May, Jun, July, Aug, Sept, Oct, Nov, and Dec. Results, the month of March, June, Aug, Sept, and Oct have P-value less than 0.05 which means these months have significant impact on the UG gas pipe damage. In addition, by avoiding digging UG gas pipe damages during March, June, Aug, Sept, and Oct when is possible will have positive impact on the UG Gas pipe damage.

**Seasons,** this attribute was split into four seasons, Autumn, Spring, Summer, And Winter. As per the UG gas pipe data distribution, damages were happening across the four seasons. Statistics analysis was performed for each season separate. Results, P-Value for Autumn, and Spring was less than 0.05, and for Summer, and Winter was more than 0.05. Which means digging during Autumn, and Spring could cause more damage to UG gas pipe than digging in Summer, and Winter. More detailed analysis included under Month Section above.

**Counties**, there are 21 counties included in the research, and statistics analysis was performed on all of the counties. For instance, County 5 has P-value equal to 0.001, and Odd ratio of 7.2. Which means county 5 has significant impact on the UG gas pipe damage, and any Gas pipe digging in county 5 has 7 times chance of been damage as compared to county 10 ( reference county). All over all, 9 counties out of 21 counties total have P – value less than 0.05. As result, UG gas pipe digging in theses counties have significant impact on the gas pipe damage, and needed to be taken in consideration when On Call 811 digging request received.

**D10, D20, D30,** these attributes was split into four ranges divided equally in order to be able to perform the statistics analysis. Results, damages within 10 miles diameter have the highest impact as compared to D20, and D30. More specifically, P value for D10 was equal to 0.001 which is less than 0.005. In addition, D20 has P value of 0.04; D30 has P Value of 0.264. This means when the agency receive digging requests, they should look at the number of damages within 10 miles diameter, and 20 miles diameter to determine the chance of the UG gas pipe been damage. However, as per this study damages within 30 miles have no impact on the UG gas pipe damage. In addition, by analyzing the effect of ranges per diameter D10, D20, and D30 on the UG gas pipe damage, the percentage within damage was higher on Range 1, than Range 2. And percentage of damage was higher on Range 2 than Range 3. And percentage of damage in Range 3 was higher than Range 4.