Questions

1.) In order to assess which portfolio is more vulnerable to winter damage, we must take into account a few factors. First, we can determine the overall risk of a portfolio by determining the total average annual loss. The total average annual loss in portfolio 1 stands at $302,941,404, and the total average annual loss in portfolio 2 stands at $557,347,527. However, care must be taken when evaluating these numbers as a portfolio that contains a significantly greater number of entries will inevitably yield higher total average annual losses, although the individual risk of each project may be less. As such, by using counters, it was confirmed that portfolio two has approximately 1.22% greater entries than portfolio one. While this number is relatively small, it's still important to confirm the average annual loss per asset in the portfolio, which will help determine which portfolio contains projects with a higher average individual risk. This can be done by taking the total average annual loss of each portfolio and dividing it by the number of projects in the portfolio. When comparing the average annual loss per asset for portfolio one and two, it is evident that portfolio one has an average annual loss per asset of $12,373.04 and portfolio two has an average annual loss per asset of $22,488.20. As such we see that the individual projects of portfolio 2, on average, have an 81.8% greater average annual loss, which indicates a significantly increased risk of winter storm damage.

2.) To conduct this analysis, I used Python to create a heatmap of correlation coefficients between different factors and the average annual loss. The correlation matrix heatmap is as follows:

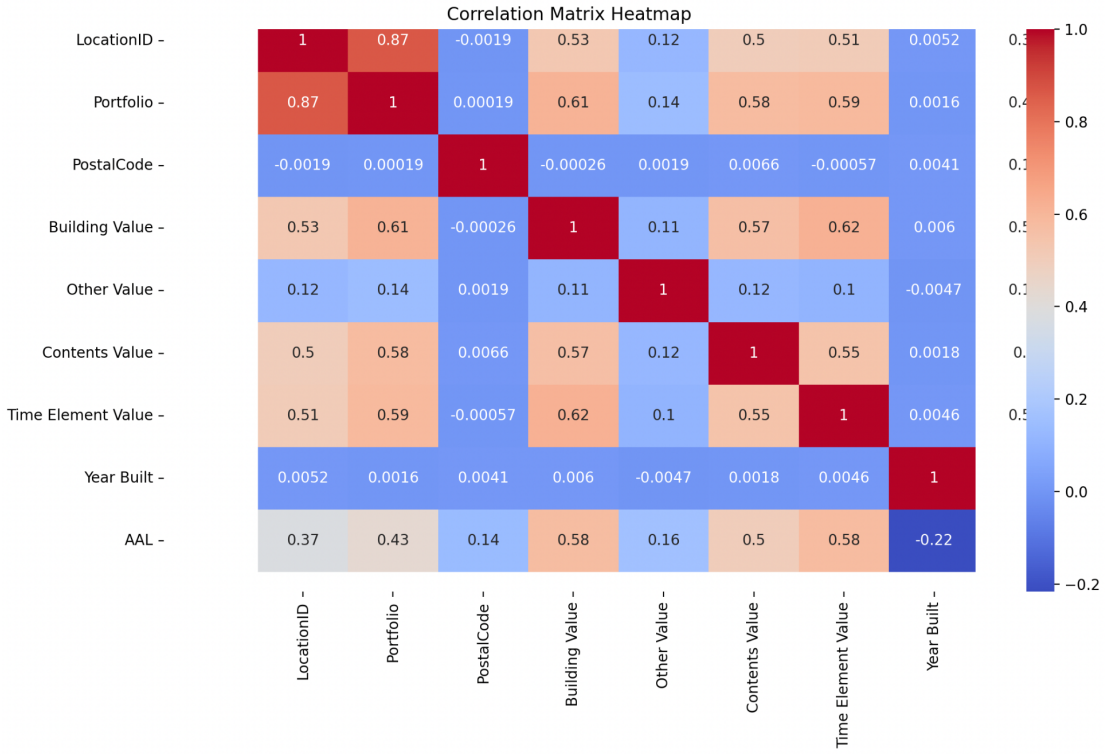


Fig 1. Shows that the Building Value and Time Element Value have the greatest association with Average Annual Loss, with a correlation coefficient of 0.58 each.

The findings show that building value and time element value have the greatest association with average annual loss, with a correlation coefficient of 0.58 each. There are a few explanations as to why the correlation of these factors is so significant. First, the building value is fundamental to determining losses as a result of weather damage, since this represents the upfront initial investment into the property. A higher investment value can indicate that the materials needed to replace the structure may be more expensive due to higher quality, or the quantity of these materials may be greater if the size of the structure is larger. Additionally, due to inflation and other factors such as value appreciation, the cost to replace damage may also increase over time.

The time element value has an equally significant correlation as building value. This is likely due to the fact a high time element value indicates loss of revenue or lost income as a result of a longer time period to restore the building for functional purposes. As such, the longer this process takes, the longer the insurance will need to cover losses incurred by the client business, significantly increasing the losses over time.

Task Assessment:

The data analysis for this assessment was solely conducted in Python. In this section of the review, I will detail explanations for how I solved each problem, cleaned the data and also explain the results. The data analysis was conducted row by row using an overarching for loop which simultaneously progressively solved all the tasks in each iteration of the loop.

1.) This task required two findings: the total insured value of each state, and the number of risks associated with each state.

In order to calculate the total insured value of an individual project the following formulas were used:

$$Total\ Insured\ Value\ (TIV) = Building\ Value + Other\ Value + Contents\ Value + Time\ Element\ Value$$

$$TIV\ of\ a\ state = \Sigma\, TIV,\ i = 1\ to\ n\ where\ n\ is\ the\ number\ of\ times\ a\ state\ appears\ in\ the\ data\ set$$

Since each state can appear multiple times, a dictionary was used to keep a track of the total insured value of the state. Each time a state was encountered, it was used as a key to determine if it existed within the dictionary. If it did, the total insured value of that project was added to the total insured value of the state. If it did not exist, then a new key-value pair was created with the state being the key and the value being the total insured value of the single row.

To calculate the number of risks associated with each state, the number of times the state appears in the data set was equal to the number of risks associated with the state. In this regard, a second dictionary was created with the state being the key and the value being the number of times that state appeared in an iteration of the data set. If the state did not exist within the dictionary it was added to the dictionary with a risk count of 1. If the state did exist within the dictionary then the risk count, which was the value, was increased by 1.

Here are two graphs that illustrate the findings. These also exist in the workbook showing the data for each question.

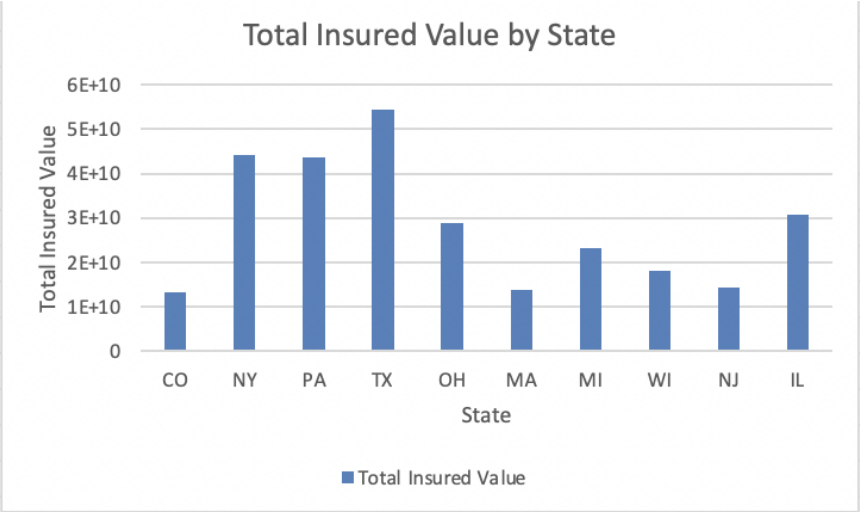


Fig 2. Illustrates the Total Insured Value per State

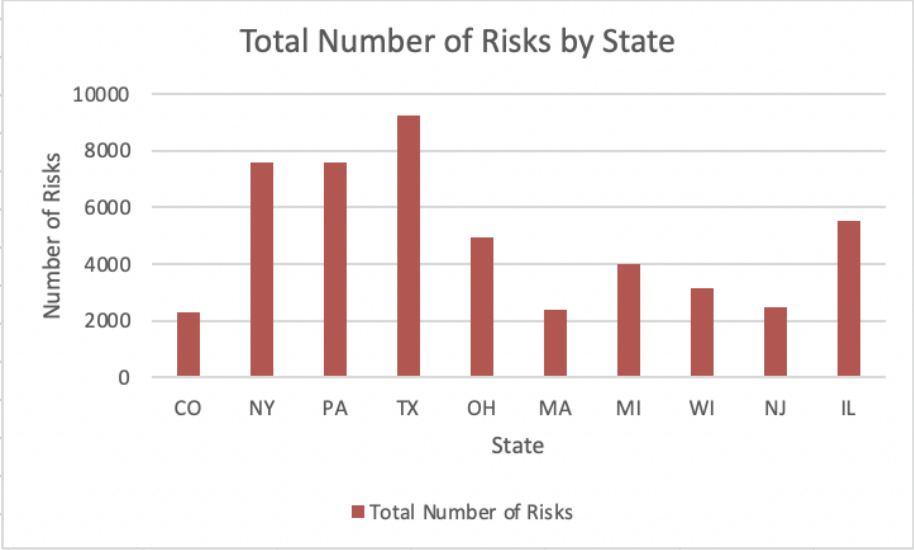


Fig 3. Illustrates the Total Number of Risks per State

A correlation between the number of risks per state and the total insured value was found. Here we see a $R^2\ correlation$ of 0.999 which indicates a significant correlation of the data. This makes sense, given that this data is regarding a solitary asset class. However, care must be taken when deriving any conclusions as a smaller number of risks with extremely high insurance values, or vice versa could theoretically disprove this correlation. This however, was not seen in the data presented, presumably due to a large sample size.

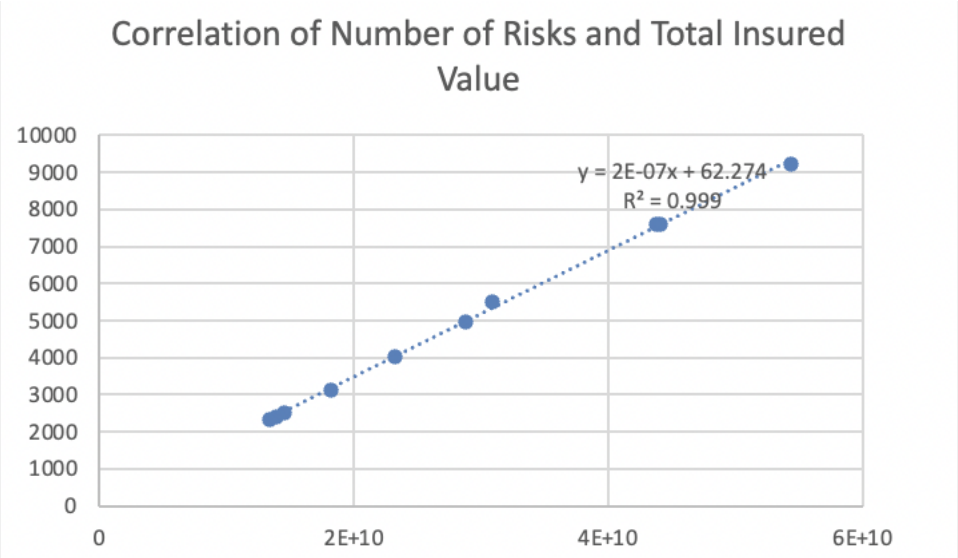


Fig 4. Demonstrates a strong correlation between the number of risks and total insured value.

2.) In order to find the five counties with the largest total insured values with construction code ‘WD10’, I similarly used a dictionary to group counties with largest total insured values if the project had the construction code ‘WD10’. If the county already existed in the dictionary, the total insured value of the project was added onto the total insured value of the county into the dictionary. Otherwise, if the county was not found in the dictionary, the county was added as a key, and the value was the total insured value of the first instance of the county in the data set.

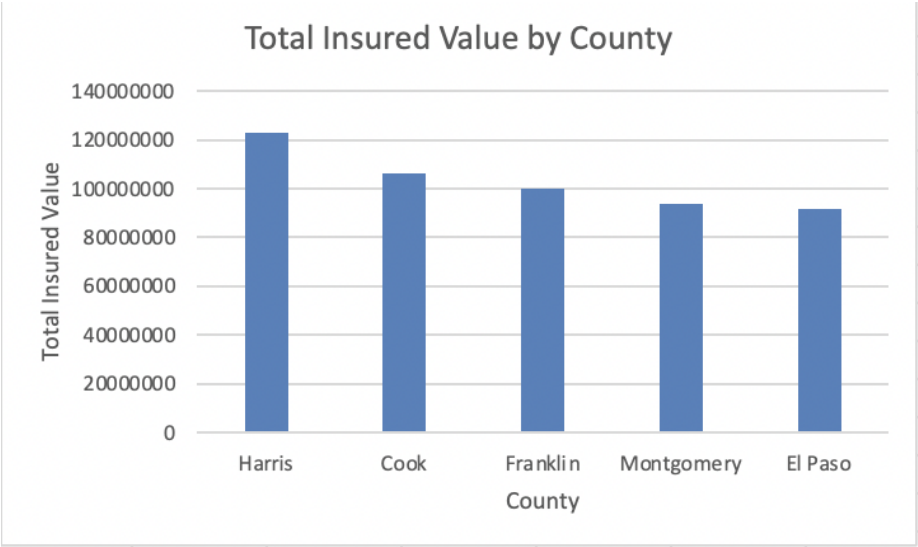


Fig 4. Illustrates the five counties with the largest total insured value with construction code 'WD10'.

3.) This task required finding the 10 postal codes most susceptible to winter damage in each portfolio. In this case, it makes sense to use average annual loss as the metric to determine susceptibility, since this is a concrete way to see financial loss over a long period of time. In order to get this data, two dictionaries were used, one for each portfolio mapping postal codes to the total average annual loss of the postal code. This was done by continuously summing the value with individual average annual losses of specific projects. Next, both dictionaries were sorted by value in descending order and stripped of all data except the 10 largest data points.

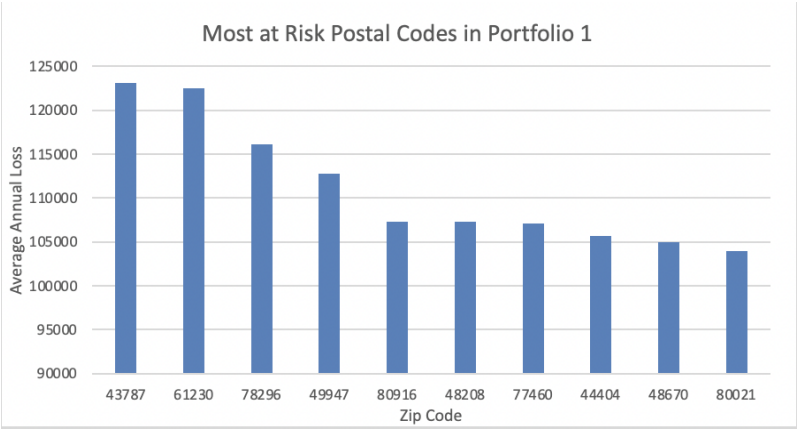


Fig 5. Illustrates the most at risk postal codes in portfolio 1.

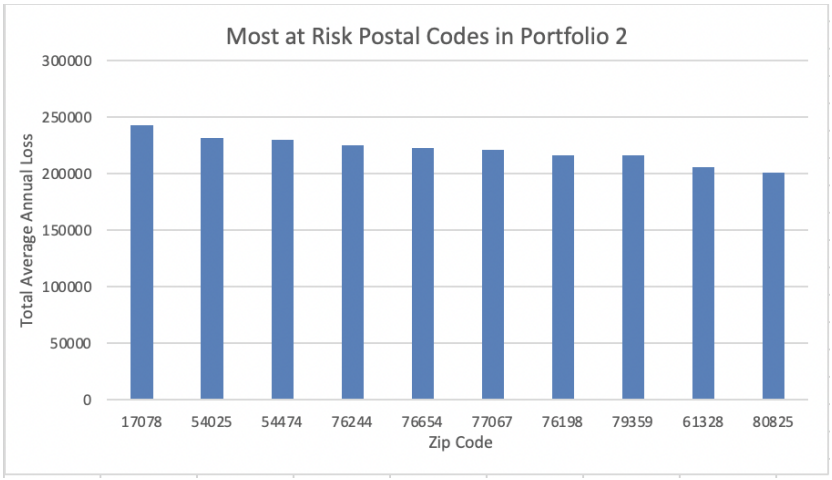


Fig 6. Illustrates the most at risk postal codes in portfolio 2.

However, this data does not necessarily give the full picture. In order to precisely examine which postal codes in each portfolio are most at risk, it is also important to consider what the average annual risk per asset is. On this account, a second set of dictionaries were created to track the risk count of zip codes in each portfolio. Next, the average annual loss of each zip code was divided by the risk count of the zip code to give an average annual loss per asset. The data is presented as follows:

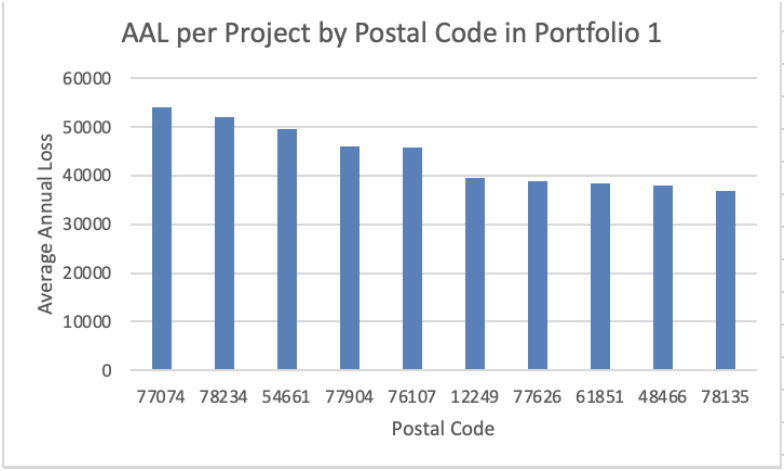


Fig 7. Illustrates the most at risk postal codes in portfolio 1 by average annual loss (AAL) per project

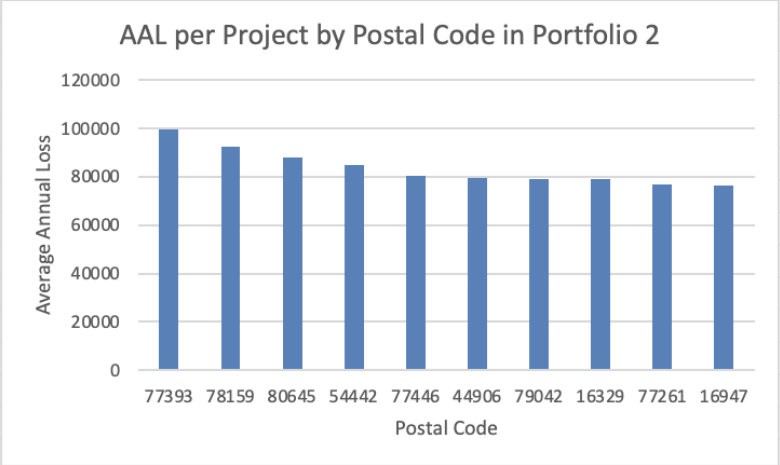


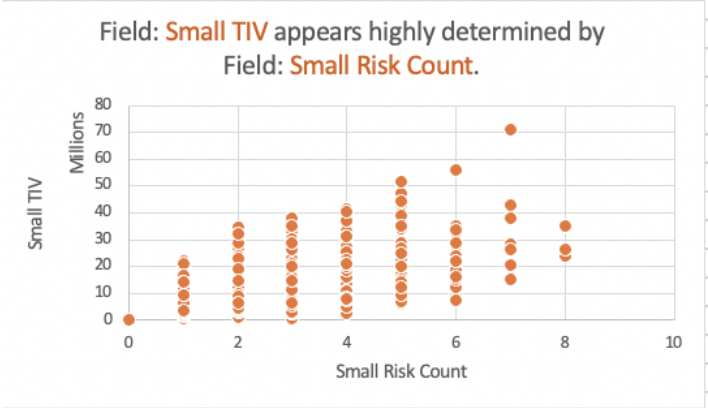
Fig 8. Illustrates the most at risk postal codes in portfolio 2 by average annual loss (AAL) per project

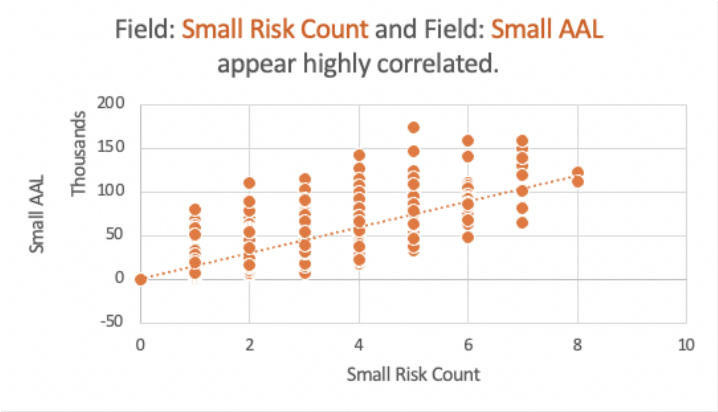
There is evidence to suggest a clear mismatch between zip codes with the highest total average annual loss and zip codes with the highest average annual loss per project in each portfolio. This is an important consideration to consider as susceptibility to winter damage should also scrutinize the environment in a zip code by looking at risk at an individual level.

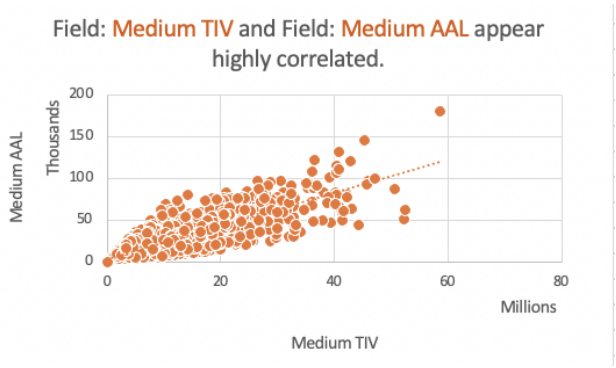
Nevertheless, the average annual losses of the largest postal codes in portfolio 2 were significantly higher than in portfolio 1.This is likely due to a significantly higher average building value in portfolio 2, which is approximately $3,889,700, in comparison to portfolio 1 which has a average building value of $573,200. This was previously proven to have a significant correlation with average annual loss with a correlation coefficient of 0.58. Similarly, the average time element value in portfolio 1 is approximately $12,373 while the average time element value in portfolio 2 is $22,488, which also had a correlation coefficient of 0.58 with average annual loss. This also suggests that winter damage in portfolio 2 is much more likely to lead to operational deficits and destruction of the core foundations of property, which further cements the concept that postal codes in portfolio 2 are larger, more capital intensive and damage will lead to prolonged repairs and construction.

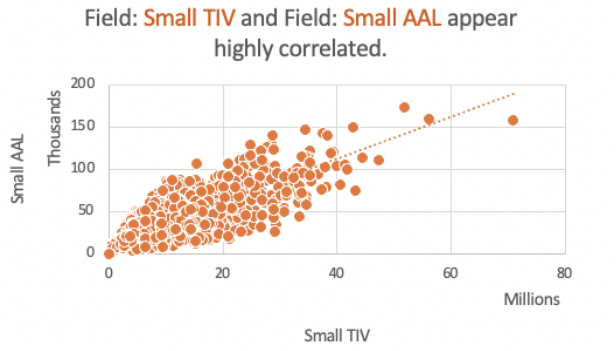
4.) The last portion of the data analysis conducted was a breakdown of total insurance value, risk count and average annual loss by building height (small, medium or large) in Pennsylvania postal codes. For each iteration of a row in the data set, first it was confirmed that the state in question corresponded to Pennsylvania. Next, the building was categorized based on the guidelines given (small: between 1 and 3 stories, medium: between 4 and 7 stories, large: greater than 8 stories). There were some entries that did not contain any data on the size of the building. For these, a separate section known as “Unknown” was used, to ensure no data was lost in the final analysis. Furthermore, since some of the data entries contained characters other than digits such as when the building was greater than 10 stories, all non-digit characters were removed and the remaining digits were subsequently converted from strings to integers. A dictionary was

declared that mapped the postal codes to the size of the buildings, each of which contained dictionaries that mapped the sizes of the buildings to Total Insured Value, Risk Count, and Average Annual Loss. While modeling this data proved to be more cumbersome due to the sheer volume, it is clear that there are some correlations to be aware of. These can be summarized clearly with the following charts:









This data analysis makes sense, as we previously proved that building value and risk count are both highly correlated with average annual loss. This large data set helped prove this for the state of Pennsylvania.