

Leveraging Synthetic Data for Maximum Predictive Power

By: Obai Kamara and Taylor Fountain



2025 ICEAA Professional Development & Training

Workshop

February 18th, 2025



Abstract

In the field of cost estimation, limited data sets often constrain the effectiveness of analytical techniques leading to less accurate predictions. This paper explores the application of Generative Adversarial Networks (GANs) for generating synthetic data to enhance cost estimation practices in scenarios with limited data. By examining the implications of synthetic data generation and the accuracy of Cost Estimating Relationships (CERs) derived from synthetic versus real data, we assess GANs' potential to improve reliability in cost estimates. This presentation will provide an overview of GANs as a machine learning technique, describe potential applications for cost estimators, identify limitations of the technique, and make recommendations for near term uses. Additionally, this paper will discuss opportunities to generate valuable insights while preserving confidentiality in data sensitive environments. The findings aim to illustrate the viability of synthetic data as a complement to traditional sources, ultimately contributing to more robust and adaptable cost estimation methodologies.



Table of Contents

| | |
|--|----|
| Leveraging Synthetic Data for Maximum Predictive Power | 6 |
| 1. Introduction..... | 6 |
| 1.1. Generative Adversarial Networks – Overview..... | 6 |
| 2. Methodology | 7 |
| 3. Analysis and Interpretation..... | 10 |
| 3.1. Exploratory Data Analysis and Preprocessing..... | 10 |
| 3.2. Training the CTGAN | 13 |
| 3.3. Synthetic Data Quality..... | 14 |
| 3.4. Regression Analysis | 19 |
| 4. Discussion..... | 23 |
| 5. Citations..... | 24 |
| 6. Appendix A – Data Set..... | 25 |
| 7. Appendix B – CTGAN Bivariate Distributions..... | 28 |



Table of Figures

Figure 1: CTGAN Model Overview. Where D = Discrete column, G = data synthesizer, T=table 8

Figure 2: Continuous Features vs. Price..... 11

Figure 3: Cut vs. Price where Carat ≈ 1 11

Figure 4: Color vs. Price where Carat ≈ 1 12

Figure 5: Clarity vs. Price where Carat ≈ 1 12

Figure 6: Correlation in Diamonds Dataset 13

Figure 7: Marginal Distributions of 'price' Feature 14

Figure 8: Marginal Distributions of 'carat' Feature 15

Figure 9: Histograms of 'cut' Feature 15

Figure 10: Histograms of 'color' Feature 16

Figure 11: Histograms of 'clarity' Feature 16

Figure 12: 'carat' vs. 'price' for Real and Synthetic Data 17

Figure 13: 'clarity' vs. 'price' for Real and Synthetic Data 17

Figure 14: 'clarity' vs 'carat' for Real and Synthetic Data 18

Figure 15: 'clarity' vs. 'price' where carat ≈ 1 for Real and Synthetic Data 18

Figure 16: Residual Plots..... 20

Figure 17: QQ Plots 21

Figure 18: PP Plot for Equation 1 on Testing Data 21

Figure 19: PP Plot for Equation 2 on Testing Data 22

Figure 20: PP Plot of Equation 1 vs Equation 2 on Testing Data 22

Figure 21: Features of Original Dataset 25

Figure 22: Features of Original Dataset 26

Figure 23: Features of Original Dataset 27

Figure 24: 'cut' vs. 'carat' for Real and Synthetic Data 28

Figure 25: 'cut' vs. 'price' for Real and Synthetic Data 28

Figure 26: 'cut' vs. 'price' where 'carat' ≈ 1 for Real and Synthetic Data 28

Figure 27: 'color' vs 'carat' for Real and Synthetic Data 29

Figure 28: 'color' vs. 'price' for Real and Synthetic Data 29

Figure 29: 'color' vs. 'price' where 'carat' ≈ 1 for Real and Synthetic Data 29



Table of Tables

| | |
|---|----|
| Table 1: Synthetic data quality metrics | 9 |
| Table 2: Hyperparameters and tuning..... | 13 |
| Table 3: CER Goodness of Fit Metrics..... | 20 |

Table of Equations

| | |
|---|----|
| Equation 1: CER Trained on Real Dataset | 19 |
| Equation 2: CER Trained on Synthetic Dataset..... | 19 |

Leveraging Synthetic Data for Maximum Predictive Power

1. Introduction

Cost estimating is a data-intensive field, yet a significant barrier to developing accurate cost estimates is the limited access to high-quality data. According to the GAO Cost Estimating and Assessment Guide, “It is often not possible for the cost analyst to collect the kinds of data needed to develop cost estimating relationships (CER) and other estimating methods” (U.S. Government Accountability Office 10). Data may be missing due to privacy concerns, data fragmentation, or logistical challenges, and even when available, it is often difficult to share due to sensitivity restrictions. As a result, cost estimates are frequently based on small datasets allowing limited inferences, which can undermine their accuracy. Addressing this issue by improving data accessibility is critical for enhancing the reliability of future cost estimates.

Most research in cost estimating has focused on exploring methods for estimating with small datasets rather than ways to increase the size of the available data set itself. Approaches have ranged from leveraging traditional statistical techniques to standard machine learning methods, each with their strengths and limitations. However, advancements in artificial intelligence, specifically generative models, may offer innovative solutions. While Large Language Models (LLMs) like GPT-4 and BERT have shown promise, older generative models, such as Generative Adversarial Networks (GANs), may be better suited for cost estimating applications due to their ability to generate realistic synthetic cost data with fewer data requirements.

1.1. Generative Adversarial Networks – Overview

GANs are a type of artificial intelligence that generates synthetic data by learning the probability distribution of an original dataset. Unlike supervised models, which require large amounts of labeled data, GANs can generate high-quality data from smaller, unlabeled datasets. Using a neural network architecture involving two models, a generator and discriminator, GANs iteratively improve their output by “competing” to produce realistic data. This process, inspired by game theory, could offer an innovative way to overcome the challenges of data scarcity in cost estimating. Beyond their well-known use in creating deep fakes, GANs have been applied in fields such as medical imaging and scientific research, where generating synthetic data can reduce costs and enhance model performance.

This study aims to explore the potential of GANs to address the challenge of limited data access in cost estimating. By leveraging GANs to generate synthetic data that mirrors real-world cost data, we hope to improve the accuracy and reliability of cost estimates, even when faced with small or sensitive datasets. This approach could represent a significant

advancement in cost estimating practices, providing analysts with more robust tools for decision-making in the absence of comprehensive historical data.

2. Methodology

Parametric analysis is a critical cost estimating technique linking estimated costs to historical data points. Parametric estimates define a statistical relationship between program costs and technical characteristics. To be done correctly, these approaches rely heavily on the collection of quality historical cost and technical data. To assess the efficacy of GANs as a tool to overcome challenges in accessing quality data, we will be generating CERs using real and synthetic cost data and comparing the outcomes of these CERs to see if there is a noticeable difference.

To generate the CERs, a large dataset of historical cost data must first be identified. Our study uses a publicly available dataset from Kaggle.com, a website which operates as a community for data scientists to collaborate on machine learning challenges. The diamond dataset from Kaggle is commonly used to learn and apply various machine learning techniques and is representative of an ideal cost estimating data source. This dataset was chosen primarily because it has cost as variable, it is large enough to test various estimating methods, and it is well documented which facilitates reproducibility and further research.

Using this dataset as a foundation, a synthetic dataset is generated using Generative Adversarial Networks. Our study tests a popular open-source GAN model, which is optimized for tabular data sources. Conditional Tabular GAN (CTGAN) is a model which generates tabular synthetic data with continuous and discrete variables as commonly seen in sensitive data sources such as medical and educational records. Most GAN and Gaussian models have difficulty generating synthetic tabular data due to the challenges presented when mixing continuous and discrete data – namely multi-modal non-gaussian distributions and imbalanced categories. CTGAN leverages mode-specific normalization, training by sampling, and a modified architecture for the generator model among other improvements to overcome these challenges¹. A high-level overview of the CTGAN model is shown in Figure 1. Below:

¹ Xu, Lei, et al. 'Modeling Tabular Data Using Conditional GAN.' *arXiv [Cs.LG]*, 2019, <http://arxiv.org/abs/1907.00503>. arXiv. pp 3-6.

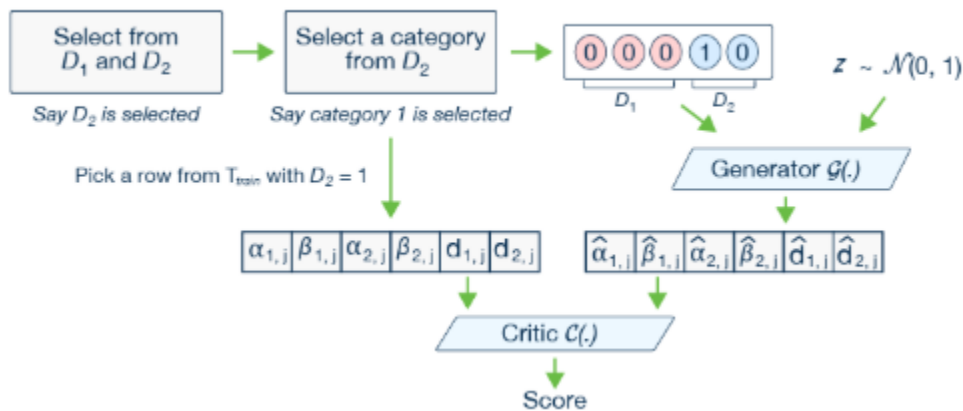


Figure 1: CTGAN Model Overview. Where D = Discrete column, G = data synthesizer, T =table

Using python, we preprocess the original diamonds dataset by filtering out datapoints with unrealistic physical measurements (e.g. a non-positive x-length, y-width, or z-depth measurement) encoding categorical fields (cut, color, clarity) to numerical values, and normalizing continuous variables – ‘depth’ and ‘table’ to percentages to make the numbers more manageable. Next, we visualize the data using a combination of box plots, correlation matrices, and scatterplots to establish initial assumptions about the relationship between the dependent and independent variables. The preprocessed dataset is then split into training, validation, and testing data sets. Using standard machine learning conventions, the training data will comprise 70% of the dataset, and it will be used to train the GAN model to learn the distribution of the data. The validation dataset (15%) will be used to tune hyperparameters, and the test dataset (15%) will assess the model’s generalization ability.

The CTGAN model provided by The Synthetic Data Vault (SDV) provides specific instructions on how to train this model. We first import the SDV library and then create a metadata file using a built-in function which identifies the features present in the dataset and classifies them as continuous or categorical. Next, we train the CTGAN on the training dataset over 500 epochs for various hyperparameters. Due to the time associated with tuning the model using a grid search strategy, the batch size was first optimized with default settings, followed by the learning rate of the generator and discriminator (which were assumed to be the same) with the tuned batch size parameter, and finally the weight decay of the underlying Adam Optimizer for the generator and discriminator (again, assumed to be the same). The ideal parameters were determined by how closely the Spearman correlation matrix of a 5000-row sample from the generator compared to the Spearman correlation matrix of the training data.

We then initialize the CTGAN model based on the training dataset, metadata, and hyperparameters and train the model to learn the distributions of the various fields present in the data. Once the model is trained, we then sample a synthetic dataset from the resulting generator.

In this paper, we assess the quality of the model, as well as the potential and shortcomings of this technique, by the quality and machine learning efficacy of the synthetic data. To assist in evaluating the quality of the synthetic data we leverage SDVs built-in evaluation functions to determine the following:

| Criteria | Metric |
|---------------------------------------|---|
| Data Validity | Continuous variables: Percentage adherence to bounds of the training data Categorical variables: Percentage adherence to the categories of the training data |
| Similarity of Marginal Distributions | Continuous variables: Complement of Kolmogorov-Smirnov statistic Categorical variables: Complement of total variance distance statistic |
| Similarity of Pair-Wise Column Trends | Two continuous variables: Normalized similarity of Spearman coefficient of the two variables: $S_{real,synthetic} = 1 - \frac{ \rho_{real} - \rho_{synthetic} }{2}$ Two categorical variables: Complement of total variance distance statistic between the normalized contingency tables of the two variables One categorical and one continuous variable: Complement of the total variance distance statistic between the normalized contingency tables of the categorical variable and the binned continuous variable |
| Anonymization | Percentage of rows in the synthetic dataset that do not appear in the training dataset Number of rows in the training dataset that appear in the synthetic dataset |

Table 1: Synthetic data quality metrics

It is important to note that scores of 100% are not necessarily the goal for the last 3 categories. Marginal and pair-wise similarity evaluating to 100% would indicate a complete duplication of the original dataset, and a particular use case may have some tolerance for re-identified datapoints. These metrics are evaluated alongside visual inspection and practicality of the datasets for a holistic assessment of data quality.

To assess the machine learning efficacy, CERs based on both real and synthetic data will be developed using generalized linear models (GLMs). While more complex models, such as decision trees or neural networks, may offer improved predictive power, GLMs were used for practicality in cost estimating, as they can provide easily interpretable and transparent results in a timely fashion. Both CERs will be tested against the same subset of the original dataset, and the CER developed on synthetic data will be evaluated both as a standalone CER and in comparison to the CER trained on real data. The process for developing these CERs will include feature selection, feature scaling, and transformation to independent and dependent variables should it be required. The CER trained on synthetic data will be evaluated on its goodness of fit to the test data, similarity to the CER trained on real data, and a comparison to the goodness of fit of the CER trained on real data.

3. Analysis and Interpretation

3.1. Exploratory Data Analysis and Preprocessing

The diamonds dataset² contains 53,941 observations of diamonds and their price, carat, cut, color, clarity, depth (total depth percentage), table (width of the top of the diamond relative to its widest point), and x, y, and z measurements (length, width, depth respectively).

The price is a continuous variable ranging between \$326 and \$18,823 dollars, and we can observe a significant positive skew. The carat, x, y, and z measurements are also continuous and positively skewed, and are observed to be highly collinear with each other. Additionally, there are several observations where x, y, and z measures are zero. The depth and table features, which both measure ratios of different measurements of a diamond, are both within the expected bounds for a percentage and exhibit low variance.

In the context of diamonds, depth can be calculated as $2*z/(x+y)$, so we remove any observations that have a volume of zero or a depth value that does not align to the established rule. This reduces the dataset to 50,017 observations.

The cut, color, and clarity features are categorical with a natural ranking of categories, making the feature a suitable candidate for ordinal encoding. It is also observed that the classes across these three features are not balanced, which may create bias in a machine learning model.

Looking at scatterplots of the continuous features against the price of a diamond (Figure 2), we observe that table and price appear to have neither a significant nor linear relationship with the dependent variable. Carat, x, y, and z do appear to have a strong relationship with the dependent variable that appears exponential. These relationships also appear similar to each other, which mirrors the hypothesized multicollinearity.

² See Appendix A for more details.

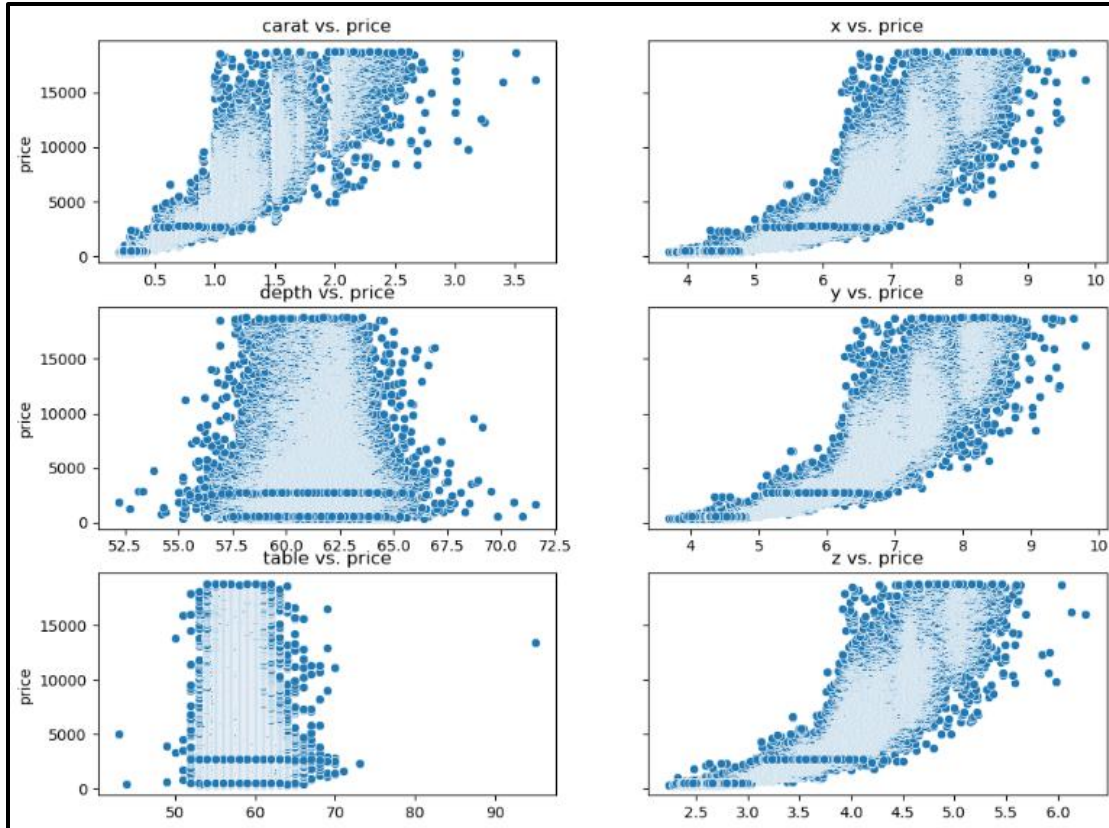


Figure 2: Continuous Features vs. Price

Looking at boxplots of the categorical features against the price of a diamond, there does not appear to be a strong correlation, much less one that aligns to the ordinal expectations of the categories. However, when controlling for carat size (Figure 3, Figure 4, Figure 5), we can see that all three categories have a positive correlation with price, albeit with varying sample and effect sizes as the carat size varies.

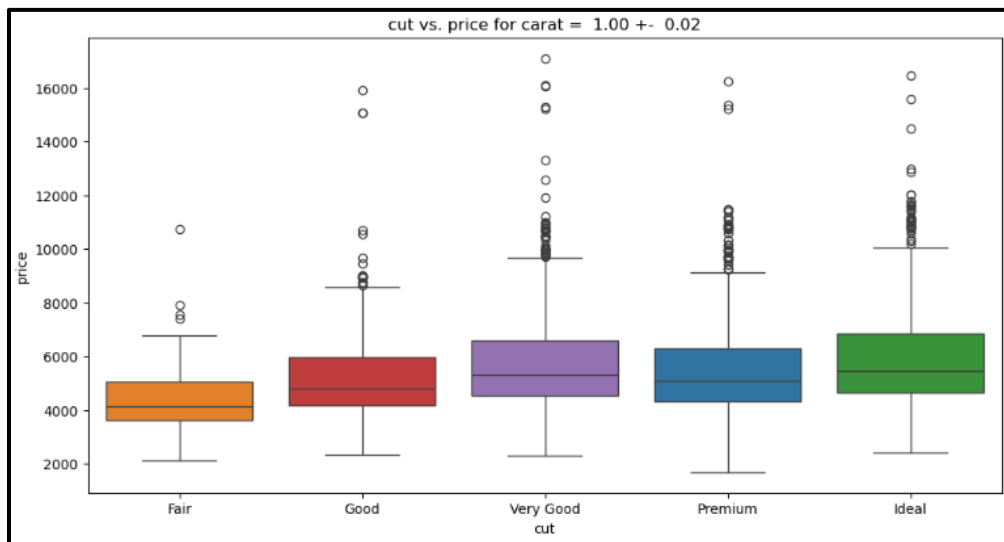


Figure 3: Cut vs. Price where Carat ≈ 1

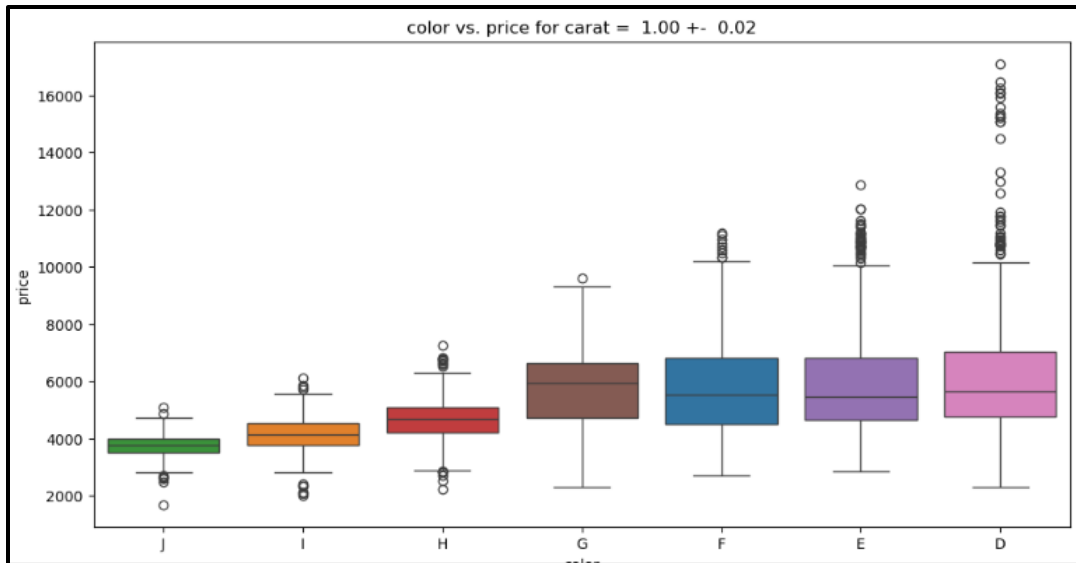


Figure 4: Color vs. Price where Carat ≈ 1

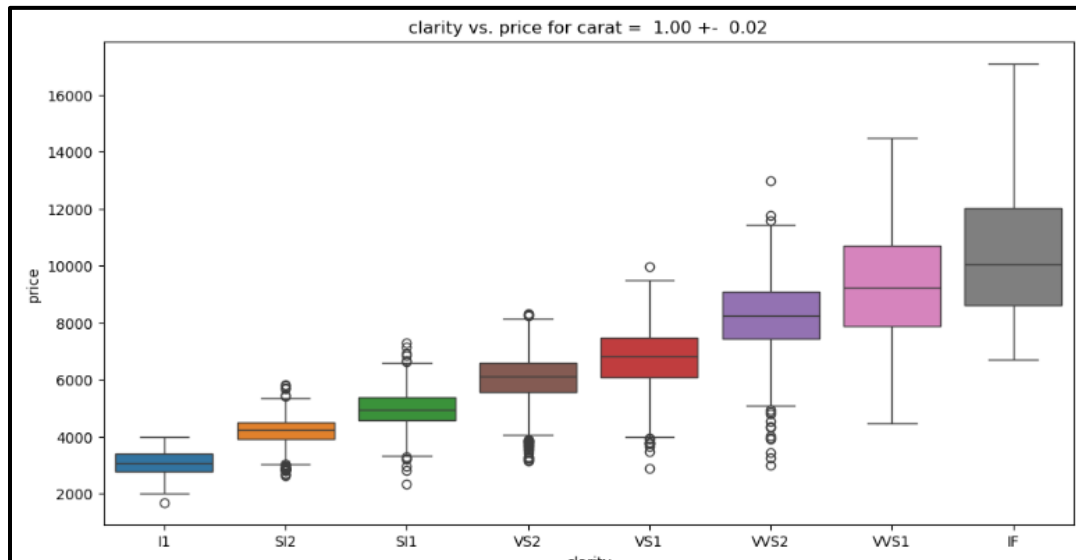


Figure 5: Clarity vs. Price where Carat ≈ 1

From the heatmap of the Spearman correlation matrix (Figure 6), we can see that our observation of strong correlations between carat, x, y, z, and price holds. To reduce bias, in both training of the GAN and training of the regression, we removed x, y, and z from the dataset. Additionally, depth and price had no correlation, and table has a minimal positive correlation to price; these features were dropped as well. We also observe that cut, color, and clarity have negative correlations to price as well as carat, the former of which does not match our observations when controlling for carat size. This indicates that while these features on their own may not have a strong relationship to price, they may have significance when interacting with the carat feature.



Figure 6: Correlation in Diamonds Dataset

The resulting dataset has 50,017 observations with five features: the dependent variable of price, the continuous independent variable of carat, and the categorical independent variables of cut, color, and clarity that are encoded with zero-based ordinal encoding. From there the data is split into training, testing, and validation subsets.

3.2. Training the CTGAN

The validation dataset was used to tune the hyperparameters of the model using the methods described in the methodology. Even when not using a grid search approach to tuning, the process took slightly under 4 hours to complete. The optimal hyperparameters are tuning process are described in Table 2.

| Hyperparameter | Tested Values | Selected Value |
|----------------|--|----------------|
| Batch Size | [1,000, 500, 250, 100] | 100 |
| Learning Rate | [1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8] | 1e-4 |
| Decay Rate | [1e-4, 1e-5, 1e-6, 1e-7, 1e-8, 1e-9] | 1e-7 |
| Epochs | Continuous; evaluated by visual inspection of loss convergence | 5000 |

Table 2: Hyperparameters and tuning

The CTGAN was then trained on the training dataset. This process took 14 hours when utilizing CPU, making the method less practical for quick turn estimates. However, this

timeline could be shortened by leveraging parallelized GPU compute and is dependent on the size of training data as well as the batch size. Using Pandas and SDV's built in CTGAN library, the process was low code to implement, making it a more accessible option than building a generative model from scratch.

3.3.Synthetic Data Quality

After training the model, 35,000 observations (about the same size as the training data) were sampled from the generator. Using SDV's built in diagnostic and data quality report functions, the synthetic dataset and training dataset were compared to evaluate the quality of the synthetic data.

All fields in the synthetic dataset were valid based on their adherence to the categories and boundaries of the training dataset. The adherence of the continuous variables was expected due to the mode-specific normalization of the method. While all 3 categorical features in the synthetic dataset adhered to the categories in the training dataset, 10 observations from the synthetic dataset had a combination of cut, color, and clarity that did not appear in the training dataset. This introduces the possibility of the model "hallucinating" or generating inaccurate results that are not realistic or possible.

Similarity of Marginal Distributions

The complement of the KS/TVD statistics for each feature ranged from 92.2-95.6%, which indicates high similarity between the marginal distributions in the real and synthetic data. Though the 2-sample Kolmogorov-Smirnov test indicates a statistically significant difference between the empirical cumulative density functions, such a difference is both expected and required for a successful CTGAN. This can even benefit in the case of masking sensitive data, so long as the practical impact is minimal in application. The high value of the complement without approaching 100% indicates the CTGAN has converged to be representative of the distribution of the training data without completely replicating it.

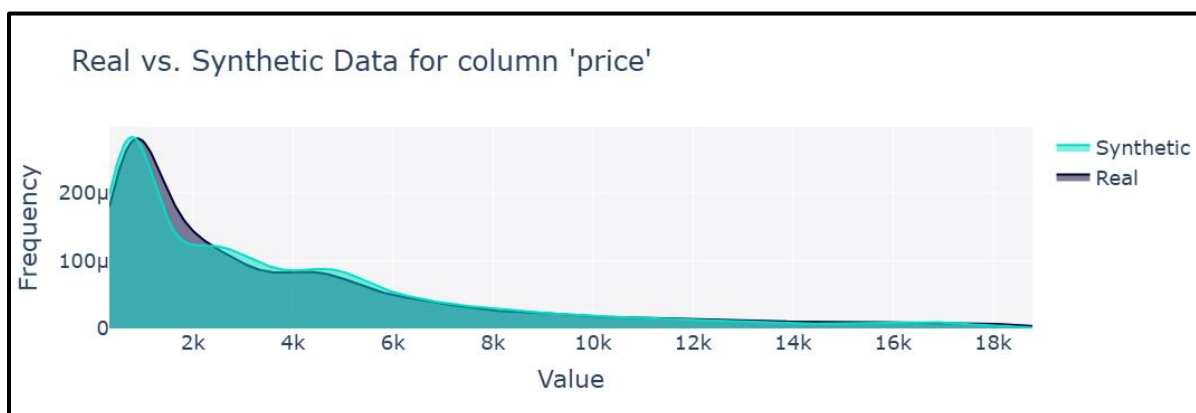


Figure 7: Marginal Distributions of 'price' Feature

By overlaying the PDFs of the real and synthetic marginal distributions for the 'carat' feature (Figure 8), we can see that the modal behavior of the real distribution is closely

replicated in the synthetic distribution, with the location of the global mode relative to all local modes maintained and the probability densities of all modes being similar in scale to the real distribution. The carat size associated with the greatest distance between the real and synthetic empirical CDFs aligns to the global mode of the synthetic data; however, this example does not make it clear if this is due to the value being the global mode or the left-most mode. The density of the right-most mode in the synthetic data has lower probability density than the right-most mode in the real data, which may reduce the leverage of outlying values (in this case, diamonds weighing approximately 2 carats) in follow-on regression analysis. It is worth noting that the vertical difference between the two PDFs is notable where the carat size is close to 0.5 and 1.25, which are both ranges between modes; this bias towards mode preservation may limit the ability to accurately predict the price of diamonds with less common carat size if there is a notable difference in the associated prices in the real data.

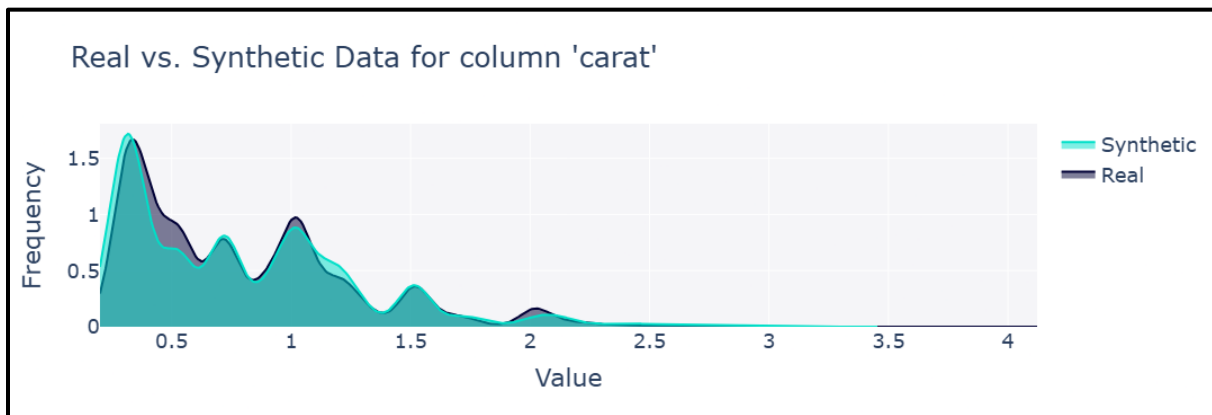


Figure 8: Marginal Distributions of 'carat' Feature

Looking at the count plots of the categorical distributions (Figure 9, Figure 10, Figure 11), we can see that the categories with the lowest number of observations in the real dataset (Fair cut, J color, and I1 clarity) are better represented in the synthetic dataset. This smoothing effect may decrease the impact of imbalanced classes in follow-on regression analysis, leading to more robust analysis of less common classes.

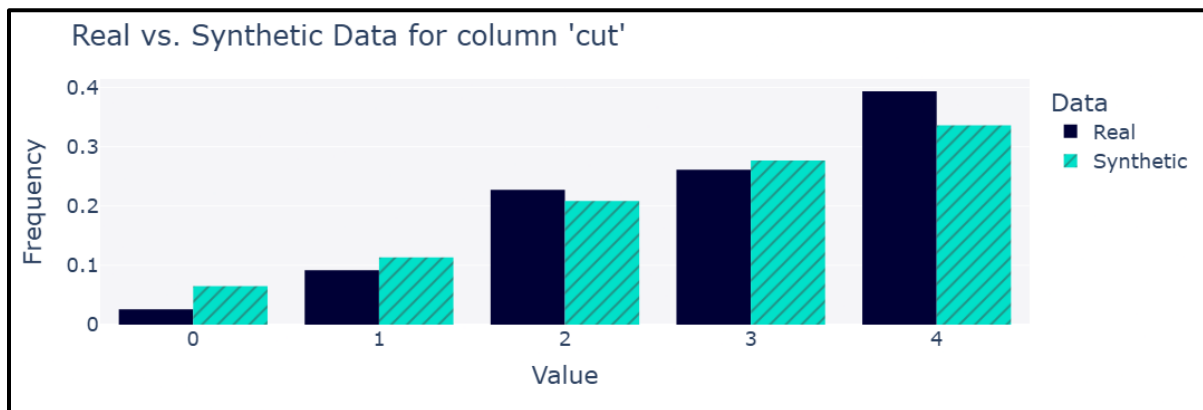


Figure 9: Histograms of 'cut' Feature

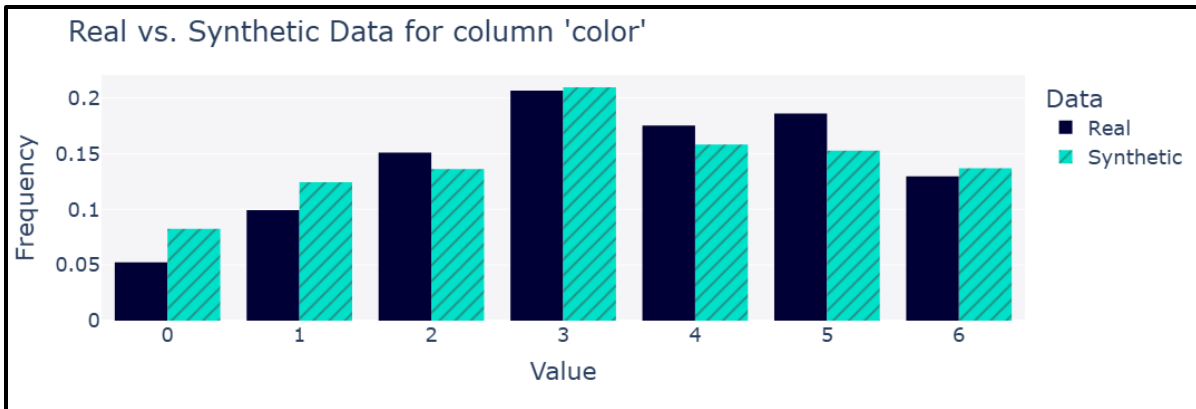


Figure 10: Histograms of 'color' Feature

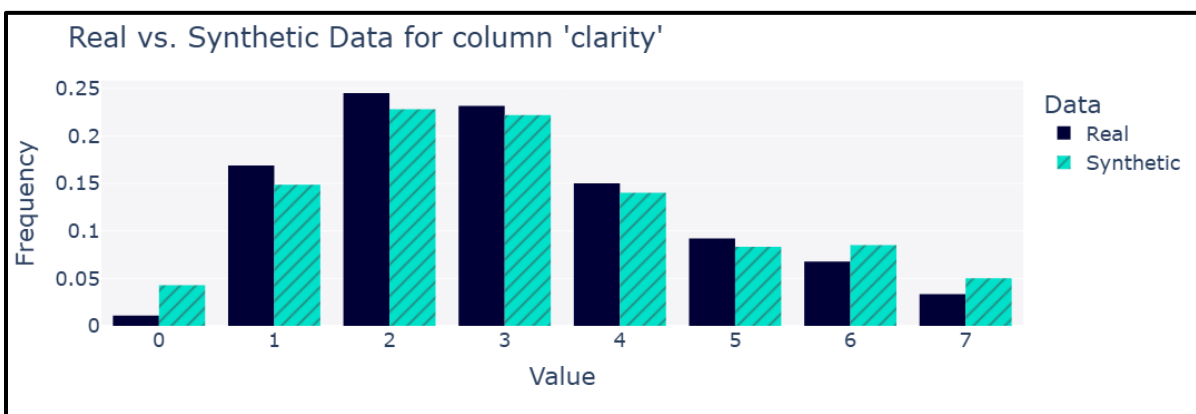


Figure 11: Histograms of 'clarity' Feature

It is notable that the KS and TVD statistics for all the independent variables range within 1.3%, with the dependent variable having a higher complement of the KS statistic. It is unknown how significant this is and could be a topic for follow-on analysis.

Similarity of Pair-Wise Column Trends

Because the goal of the synthetic data is to develop CERs, maintaining the relationships between variables is crucial. Because the strongest relationship identified in the real dataset was between carat size and price and the relationship between the three categorical features and price appears to be dependent on the carat size, maintaining this relationship is paramount.

The Spearman correlation coefficient between the carat size and price in the real data was 0.96, while the coefficient in the synthetic data was 0.92. This makes the normalized correlation similarity between the two datasets 98%, which indicates that while the trend is largely maintained in the synthetic data, there is some deviance from the real data. As some degree of deviance in the correlation coefficients is expected in synthetic data and required for a CTGAN, the existence of a difference between the two coefficients is only a problem if it has negative impacts in application. However, the high degree of similarity is a

positive indicator that the relationship between the two variables will be captured in follow-on analysis.

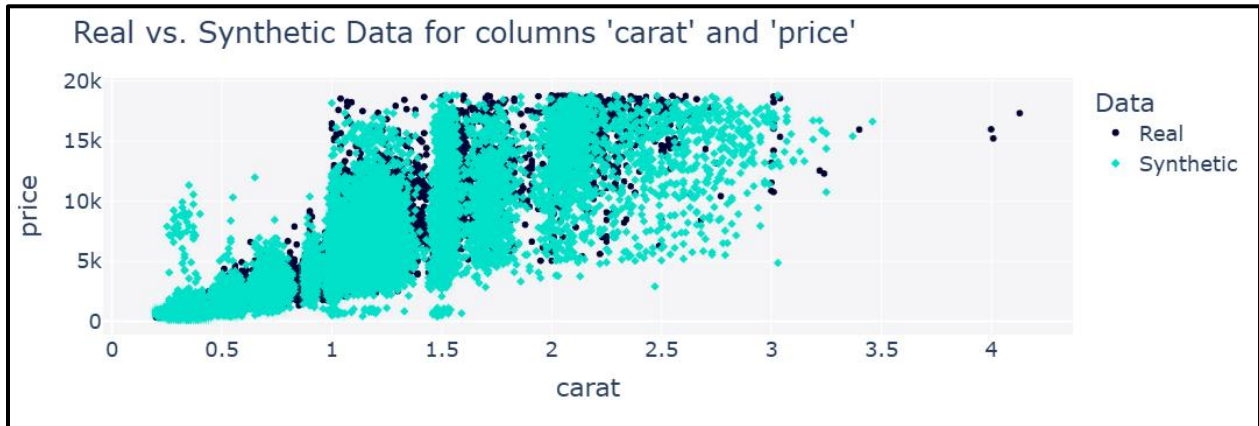


Figure 12: 'carat' vs. 'price' for Real and Synthetic Data

Looking at the overlaid scatterplots of the carat size vs. price for both datasets (Figure 12), we can see that there is a cluster of synthetic observations with carat size .3-.6 that have prices significantly higher than the real observations as well as another cluster with carat size 1-1.5 that are priced at less than \$1,000 dollars, which appear to have a notable effect on the correlation coefficient. It is typically recommended to not remove outliers from a dataset, as they are likely representative of some extreme condition that is in the realm of possibility. However, without the proof inherent in real data that extreme conditions have occurred, it is possible that outlying datapoints are not representative of an extreme condition.

Under the assumption that the CTGAN was sufficiently trained to identify points so outlying that their probability was statistically negligible, we decided to only exclude data points that were deemed physically impossible. As the features and training data were selected in such a way to prevent this, no outliers were removed from the synthetic data.

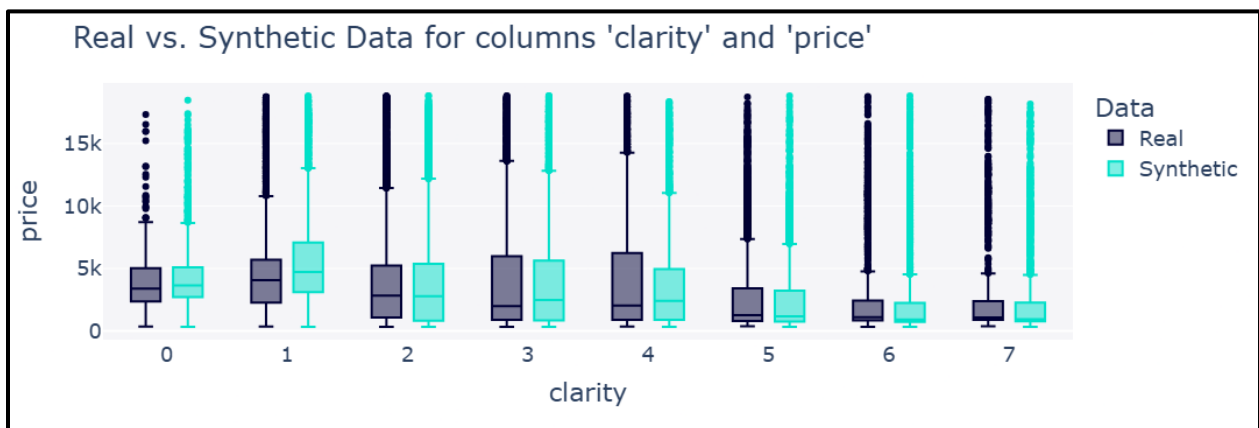


Figure 13: 'clarity' vs. 'price' for Real and Synthetic Data

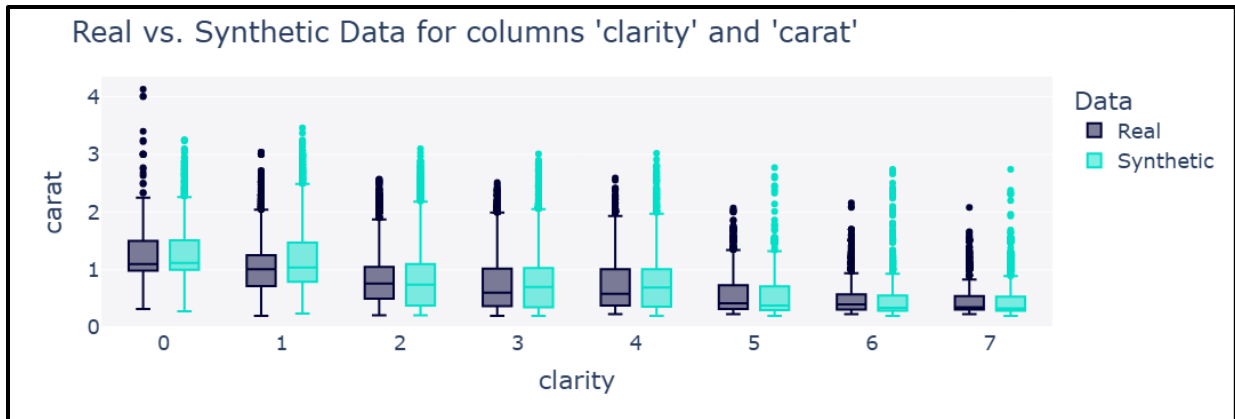


Figure 14: 'clarity' vs 'carat' for Real and Synthetic Data

We observed in EDA that the correlation between price and the categorical features were not representative of their relationship. We can observe from the boxplots that this is similarly true with the synthetic data, as well as the weak negative correlation with the categorical features and carat size (Figure 13, Figure 14). Once again controlling for carat size, we can see that the effect size of the interaction item between carat and categorical features does not appear as strong (Figure 15). It is unclear how large an effect this will have on follow-on regression analysis. For more plots, see Appendix B – CTGAN Bivariate.

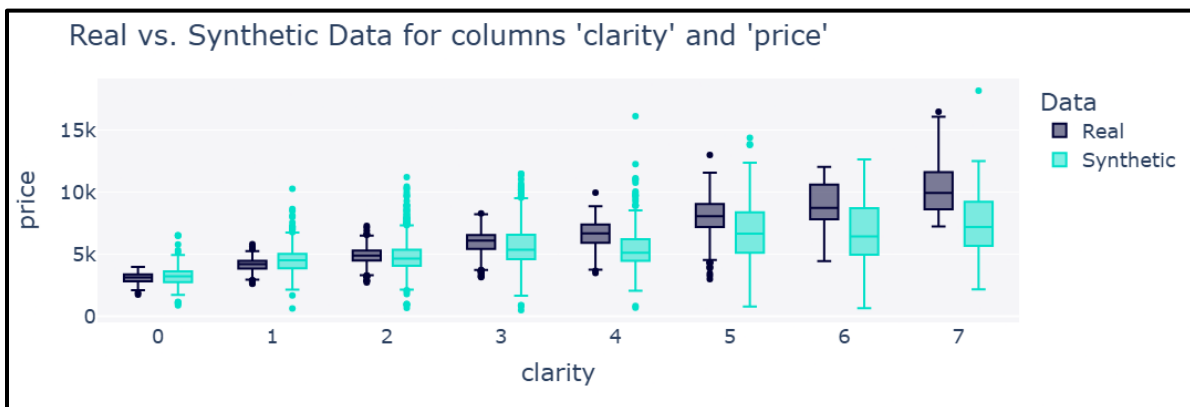


Figure 15: 'clarity' vs. 'price' where carat ≈ 1 for Real and Synthetic Data

Anonymization

160 observations are shared between the two datasets, which is .45% of the real dataset. While this may be low in some use cases, there may be highly sensitive data that has a 0% tolerance. If there is no tolerance for re-identified data points, the analyst has the option to either retrain the model or remove them from the dataset prior to distribution, which, due to the low percentage and significant sample size, would have minimal impact on the synthetic data's utility.

Various techniques also exist for contextualized anonymization, which masks personally identifiable information (PII) or other sensitive information while preserving certain

sensitive contexts, such as geographical data. While this has a larger impact in the medical field where PII is handled regularly and does not apply to the data in this experiment, it can be applied in cost estimating contexts for masking organization- or security-sensitive data.

3.4. Regression Analysis

Both the real and synthetic datasets underwent feature engineering and CERs were fitted using weighted least squares regression. Though the datasets were assessed for transformations independently, the same transformations were implemented due to their similarity.

Pre-processing and Training

Due to the significant positive skew of both the price and carat size features, both underwent Yeo-Johnson transformations to normalize the distributions. The price was then log-transformed to address the remaining non-linearity, and the independent variables were shifted to have a minimum value of 1, log-transformed, and then standardized.

The CERs were originally developed using OLS, and then the residuals were assessed for heteroskedasticity. The regression was then re-run using WLS. Various powers of price, carat, and price per carat were trialed for an appropriate weight. For the real training data, the untransformed price from the training dataset minimized errors and heteroskedasticity of the residuals for all weights tested. For the synthetic data, the carat⁻⁸ minimized errors and heteroskedasticity of the residuals for all weights tested.

Assessment of CERs

The following CER was yielded from the regression trained on the real dataset:

$$price = (1 - .2423 * (cut + 1)^{.0075} * (color + 1)^{.0249} * (clarity + 1)^{.0410} * (2.062 - (carat + 1)^{-1.268})^{1.078})^{-13.97} - 1$$

Equation 1: CER Trained on Real Dataset

The following CER was developed from the regression trained on the synthetic dataset:

$$price = (1 - .0238 * (cut + 1)^{.0051} * (color + 1)^{.0241} * (clarity + 1)^{.0431} * (1.943 - (carat + 1)^{-1.129})^{1.250})^{-193.3} - 1$$

Equation 2: CER Trained on Synthetic Dataset

Both CERs were tested on the testing dataset, with observations in the testing dataset that did not adhere to the bounds of the training data removed. The results were evaluated using goodness-of-fit metrics, diagnostic visualizations, and comparison to each other.

The goodness of fit metrics for each CER was as follows:

| Training Data | R ² | RMSE | MAPE | Mean AE | Median AE |
|---------------|----------------|----------|-------|----------|-----------|
| Real | .960 | \$808.18 | 10.9% | \$404.37 | \$165.57 |
| Synthetic | .948 | \$915.19 | 12.3% | \$476.79 | \$183.89 |

Table 3: CER Goodness of Fit Metrics

Both sets of results indicate similar performance. The R² scores for both CERs indicate a high degree of accuracy, but the RMSEs being more than twice the minimum value of the real prices indicates the presence of some large residuals. The MAPE values are both within an acceptable range, which indicates the large residuals occur with larger real prices, a conclusion that aligns with the assumption of multiplicative errors in an exponential model. This is supported by the significant difference between the mean and median absolute errors, the latter of which indicates that the results are highly accurate for more than half of the testing observations, even if the right tail behavior is not captured as well.

As is expected, the CER trained on real data performs better than the CER trained on synthetic data. However, the relative scale and relationships between the different metrics are maintained in the CER trained on synthetic data, and when assessed independently the latter CER appears to perform well in its own right.

Looking at the residual plots (**Error! Reference source not found.**), we can see that the standardized residuals of both CERs exhibit homoskedasticity with 95% of the residuals falling within two standard deviations of the mean. There is also a ceiling effect at the right tail and a handful of highly outlying residuals, which can explain the significant differences in the mean and median absolute errors in both models. The similarity of the graphs and adherence of the synthetic data to the assumptions of linear regression are both positive indicators for the utility of the synthetic dataset.

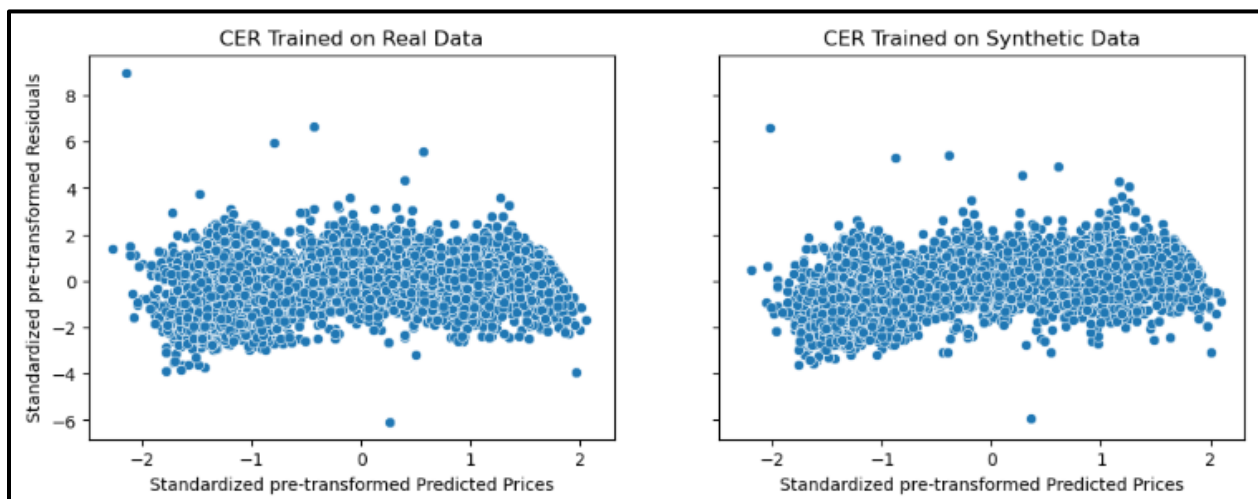


Figure 16: Residual Plots

Looking at the QQ plots of the standardized residuals of both CERs (Figure 17), we can see that the errors are normally distributed to at least 3 standard deviations. The sharp

deviation from this trend for extreme values indicates both models struggle with capturing tail behavior. The higher maximum standardized residual in the CER trained on real data indicates an increased sensitivity to outliers, which could imply the CTGAN smoothed the data, which would give less-observed points more significance, at the cost of slightly decreased accuracy of the model.

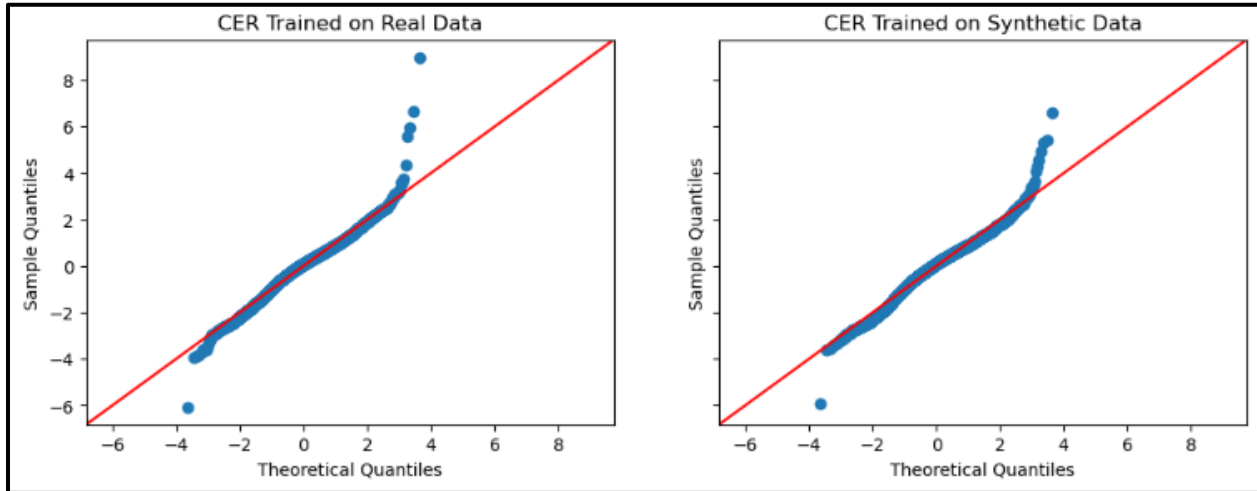


Figure 17: QQ Plots

The PP plots of both CERs compared to the actual values (Figure 18, Figure 19) are almost identical in shape, including a shelf effect at $\sim 45\%$ and a noticeable difference in behavior at the left tail. Even though this indicates sub-optimal performance of the CER, it is a positive indicator for the synthetic data capturing atypical behavior in the real data. In fact, when plotting the CDFs of the standardized results against each other (Figure 20), we can see that the shapes of the distributions are almost identical.

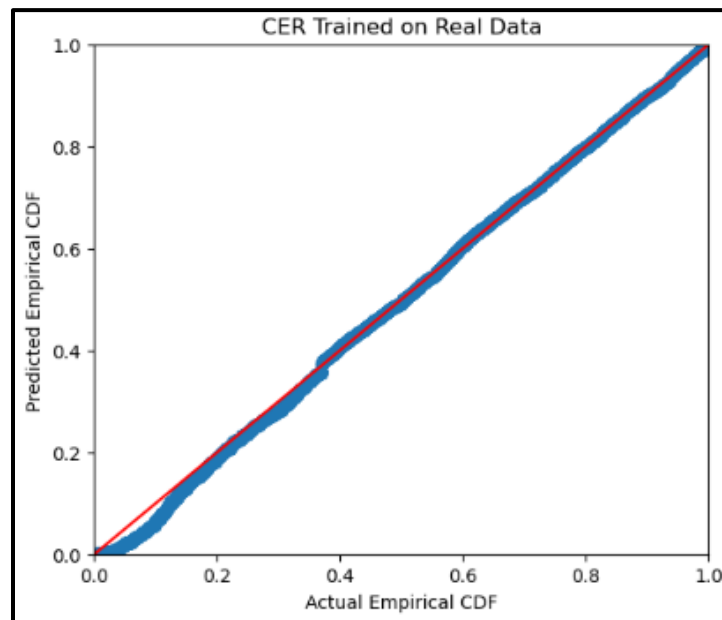


Figure 18: PP Plot for Equation 1 on Testing Data

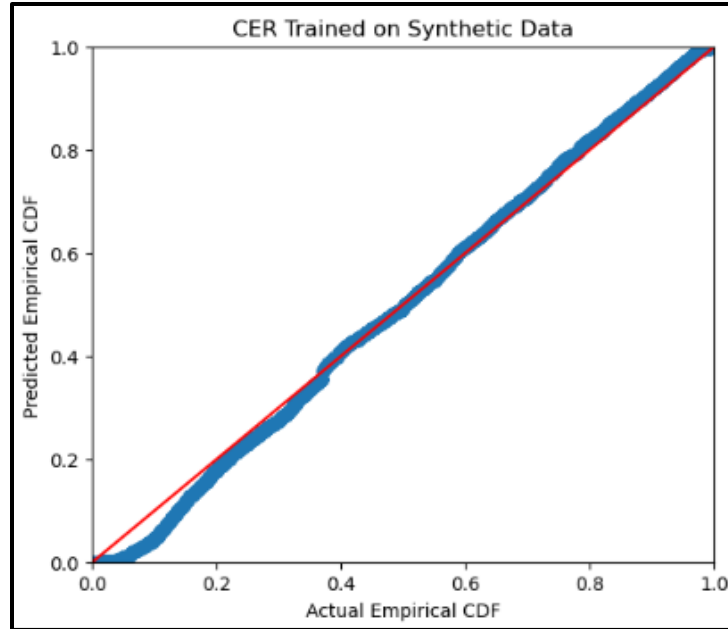


Figure 19: PP Plot for Equation 2 on Testing Data

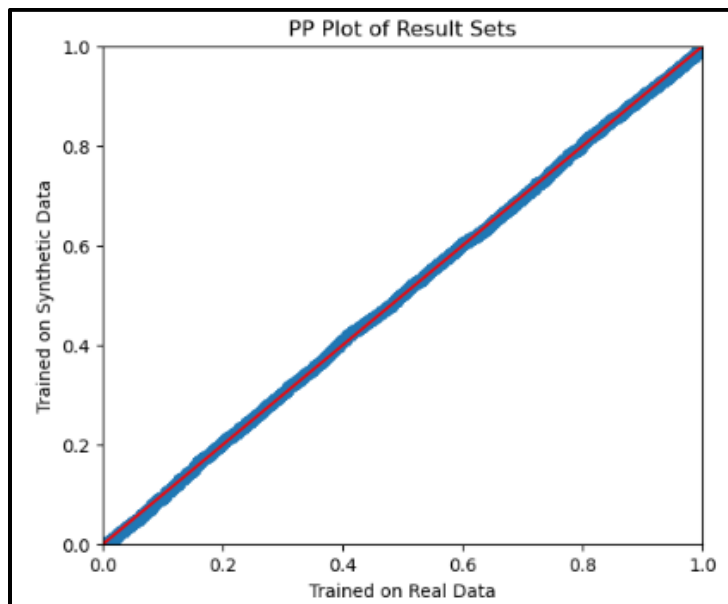


Figure 20: PP Plot of Equation 1 vs Equation 2 on Testing Data

Finally, we compare the transformed results. The maximum absolute residual for Equation 2 is smaller than Equation 1; this indicates that Equation 2 is less sensitive to outliers even if it is less accurate. However, the residuals for Equation 2 are smaller than that of Equation 1 for 38.8% of the testing dataset, which combined with the satisfactory goodness of fit metrics suggests that training on synthetic data can have a minimal impact on the utility of a CER. When looking at the difference in the predicted values between the two CERs, 50% of the predictions differ by less than 60 dollars, which shows that estimates based on Equation 2 are fairly close to that of Equation 1 the majority of the time.

4. Discussion

The quality and diagnostic scores of the trained GAN model verify that highly realistic cost datasets can be created. These synthetic datasets replicate the marginal distributions of the original dataset while also preserving the relationships present between categorical and numerical features. Cost Estimating Relationships generated using synthetic data perform well as compared to the CERs generated using the true dataset and enable analysts to identify complex relationships present in the synthetic data.

The quality of these results did come with a few limitations. Implementing the CTGAN model requires significant resources in terms of personnel understanding machine learning as well as time required to train the model. Additionally, the model is trying to represent the true dataset as best as possible, but if that true dataset has any biases (from data gathering or naturally occurring) those biases will also be replicated in the synthetic dataset. Domain knowledge is also paramount when training and evaluating the synthetic dataset as it is on the analyst to identify if an observation is out of the realm of possible i.e., the model is hallucinating. In our specific example we found that the GAN struggled to capture the relationship between price and carat between .3-.6 and 1-1.5 carats. Any significant under or over estimations can be curtailed with specific parameters during the training process. Finally, if there are observations that are exact replicates of the true dataset, the analyst needs to address how the model should be updated to preserve anonymity.

While training GANs are computationally intensive and require a highly technical skillset to implement properly, they offer opportunities to improve cost estimating as a field. The results of this research would be most beneficial for organizations, programs, and projects with large proprietary or sensitive datasets. Often, these datasets are used internally for research or analysis and only the results (e.g. a CER) are shared with the larger community. With organizational backing, sharing synthetic versions of these datasets would empower the cost estimating community to do additional research or develop alternative CERs – facilitating collaboration in a way that has been previously limited.

GANs offer a promising approach to improving cost estimation in data sensitive environments. While synthetic data does not replace the need for high quality historical data, it serves as a valuable supplement which can enhance estimating reliability while preserving data confidentiality. This paper scratches the surface of the applicability of GANs in cost estimating. Further research into GAN effectiveness on smaller datasets, testing the use of more complex CER models, and comparison to more traditional techniques such as Monte Carlo simulation would improve how these tools can be leveraged in the community moving forward.

5. Citations

| Citation | Link |
|--|---|
| Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN. NeurIPS, 2019. | https://arxiv.org/pdf/1907.00503 |
| Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In Advances in Neural Information Processing Systems, 2017 | https://arxiv.org/pdf/1705.07761 |
| Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Machine Learning for Healthcare Conference. PMLR, 2017. | https://arxiv.org/pdf/1703.06490 |
| Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. Commun. ACM 63, 11 (November 2020), 139–144. https://doi.org/10.1145/3422622 | https://dl.acm.org/doi/abs/10.1145/3422622 |
| Fedus, William, et al. ‘Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step.’ arXiv [Stat.ML], 2018, http://arxiv.org/abs/1710.08446 . arXiv. | https://arxiv.org/abs/1710.08446v3 |
| Xu, Lei, and Kalyan Veeramachaneni. ‘Synthesizing Tabular Data Using Generative Adversarial Networks.’ arXiv [Cs.LG], 2018, http://arxiv.org/abs/1811.11264 . arXiv. | https://arxiv.org/abs/1811.11264 |

6. Appendix A – Data Set

Data Source: <https://www.kaggle.com/datasets/shivam2503/diamonds>

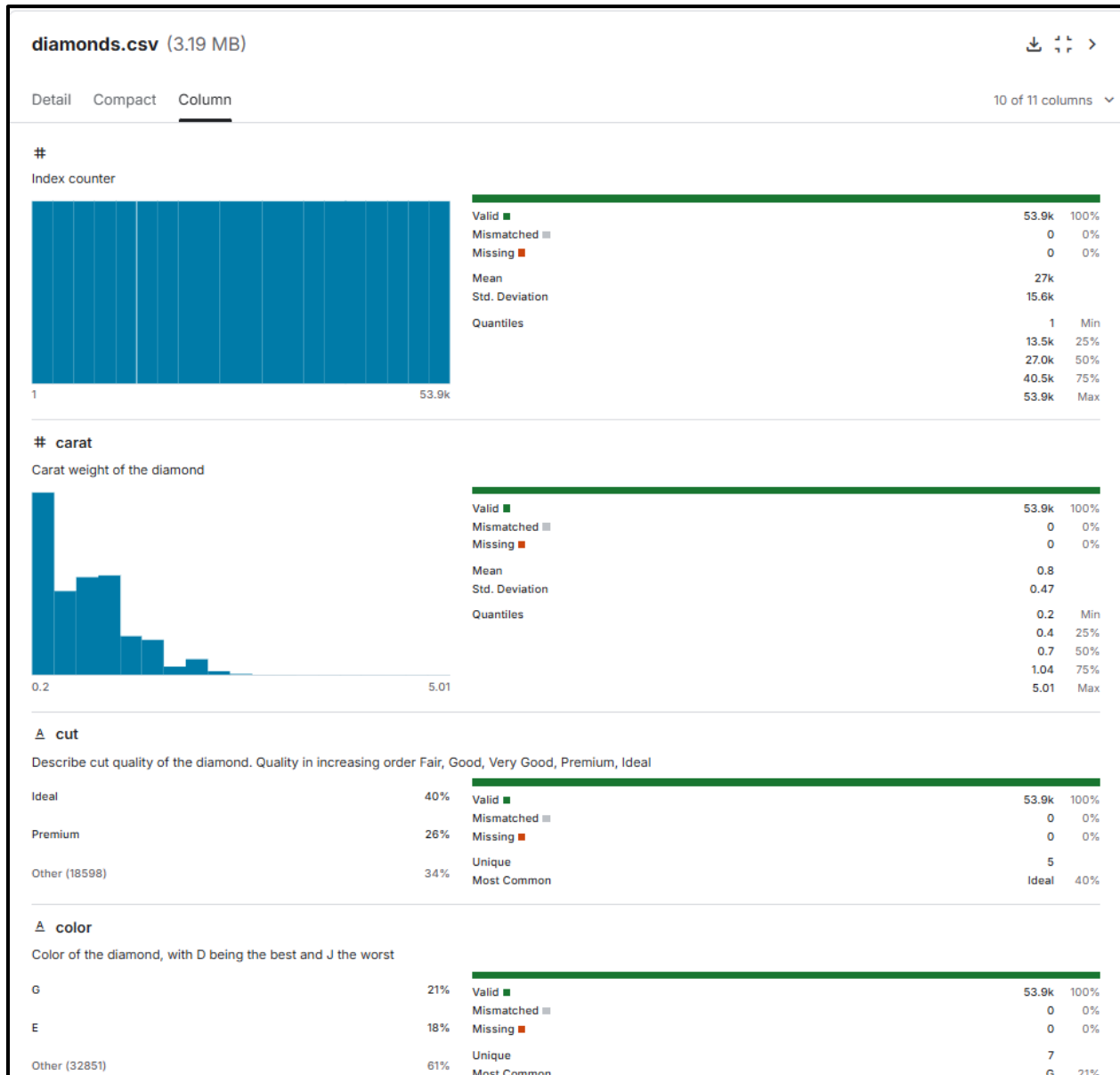


Figure 21: Features of Original Dataset

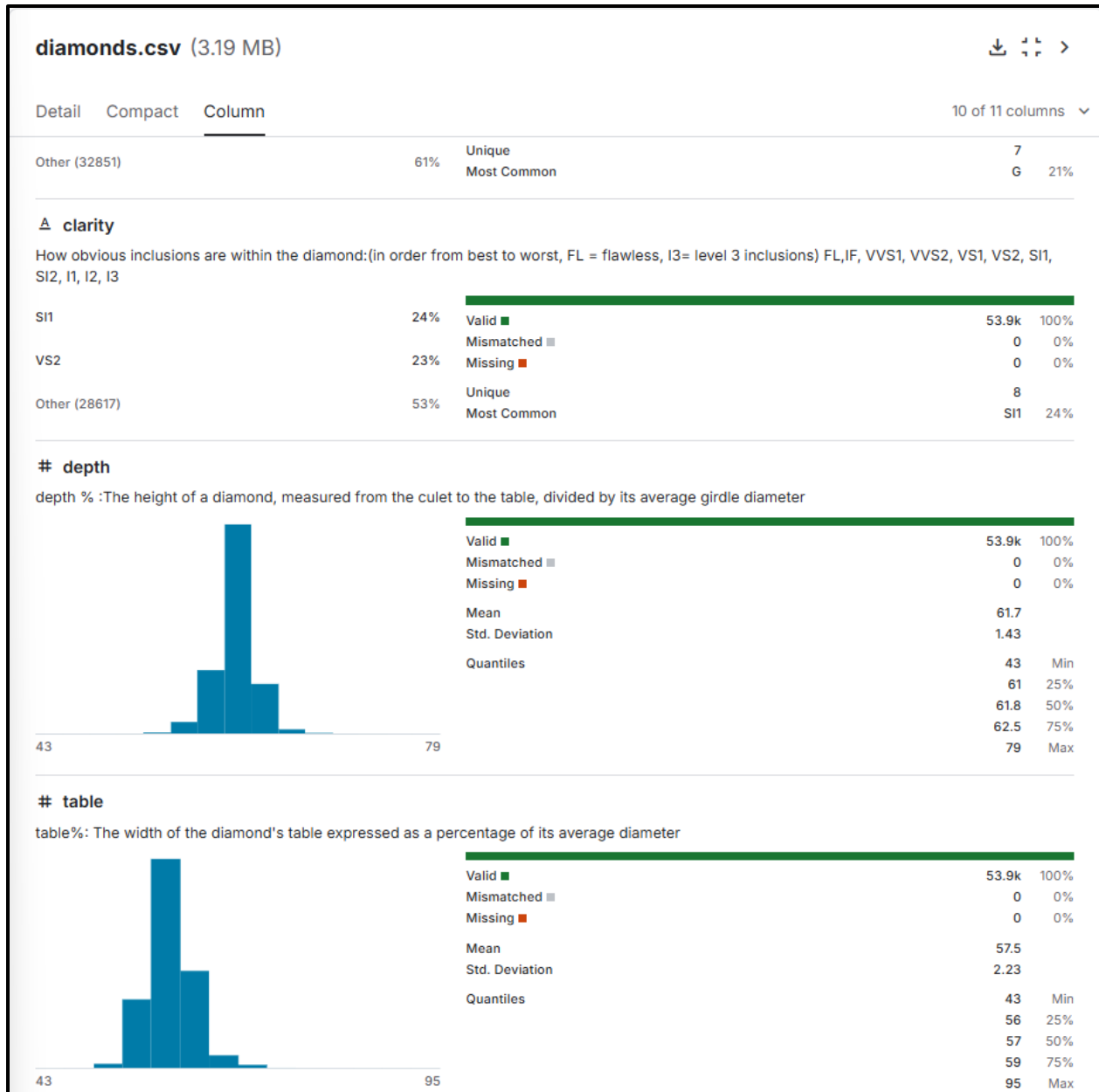


Figure 22: Features of Original Dataset

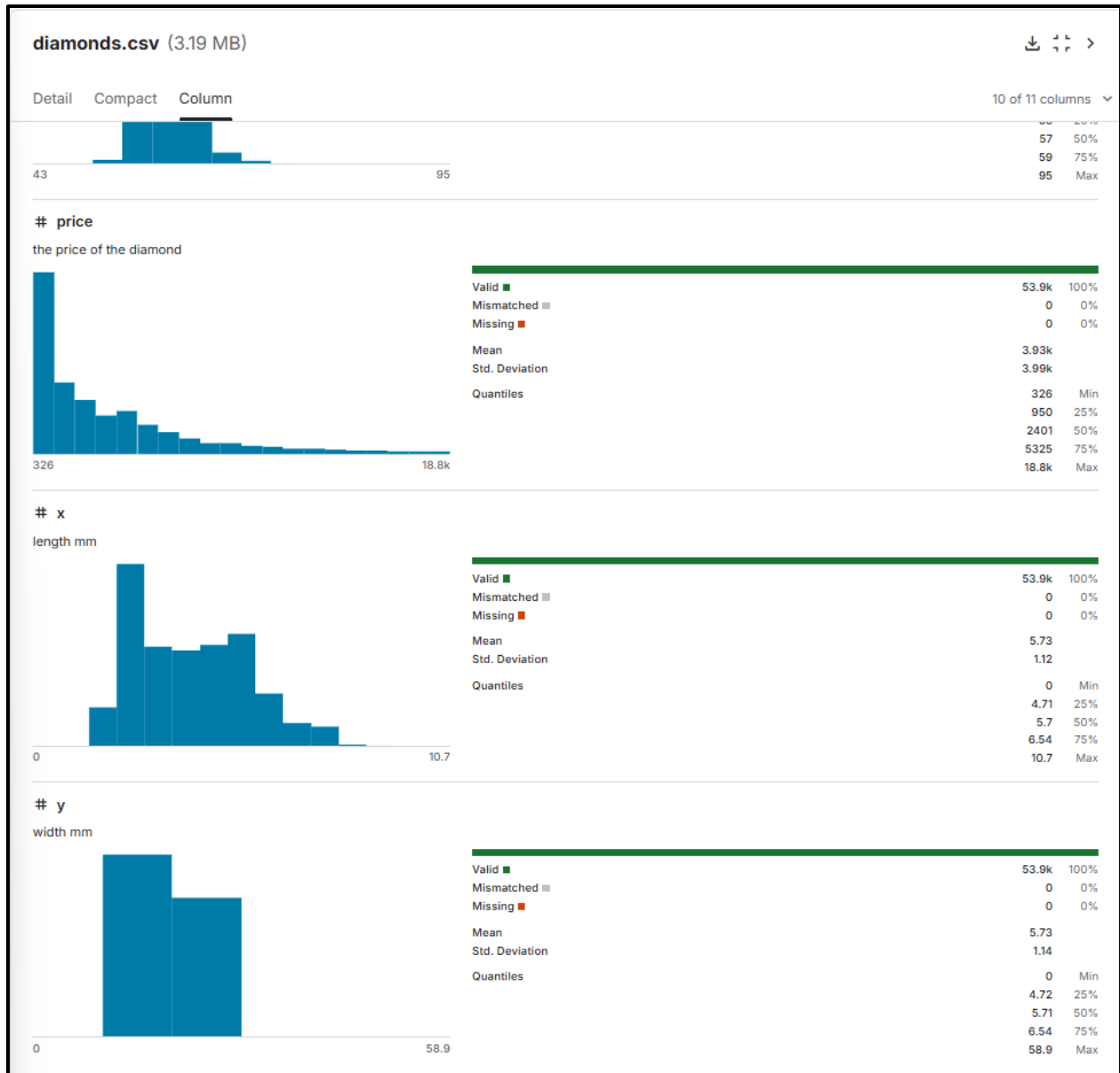


Figure 23: Features of Original Dataset

7. Appendix B – CTGAN Bivariate Distributions

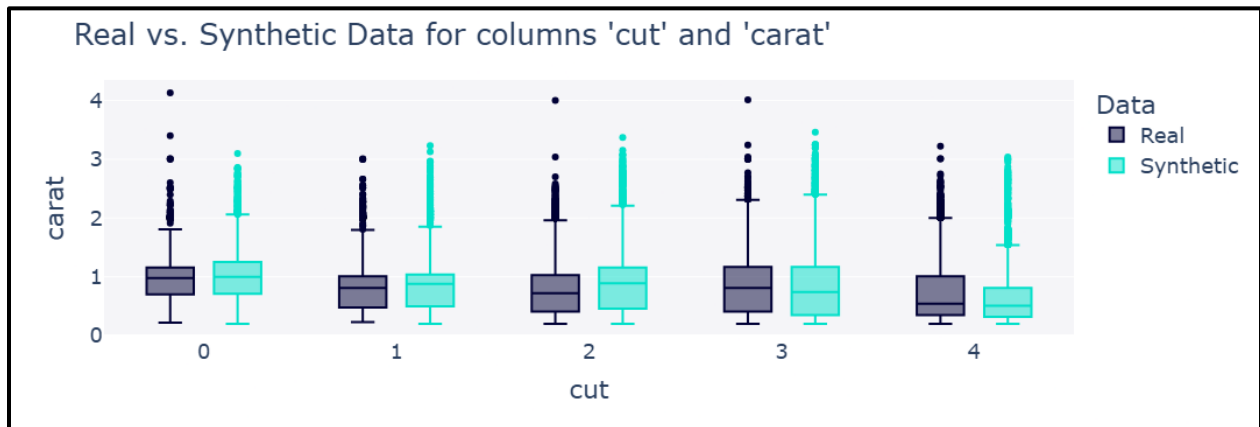


Figure 24: 'cut' vs. 'carat' for Real and Synthetic Data

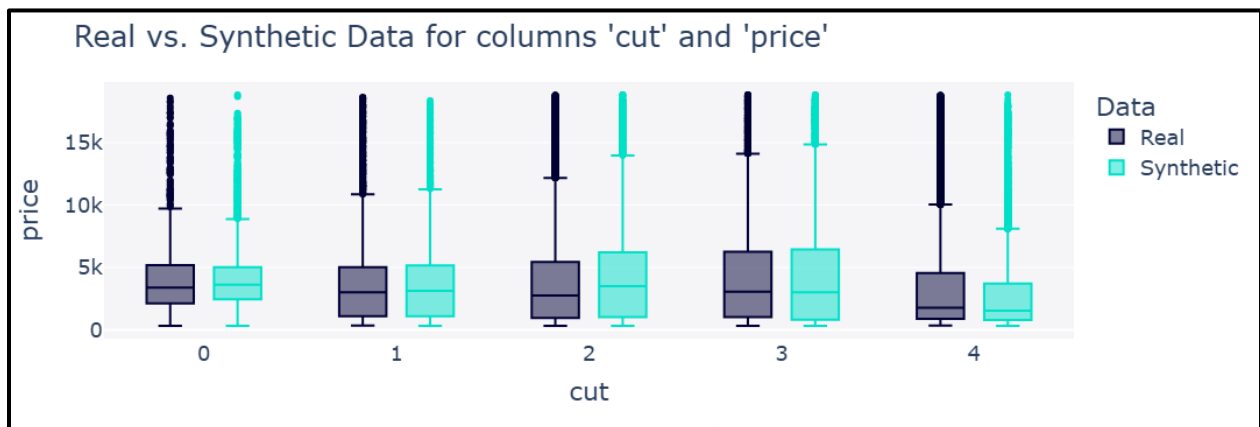


Figure 25: 'cut' vs. 'price' for Real and Synthetic Data

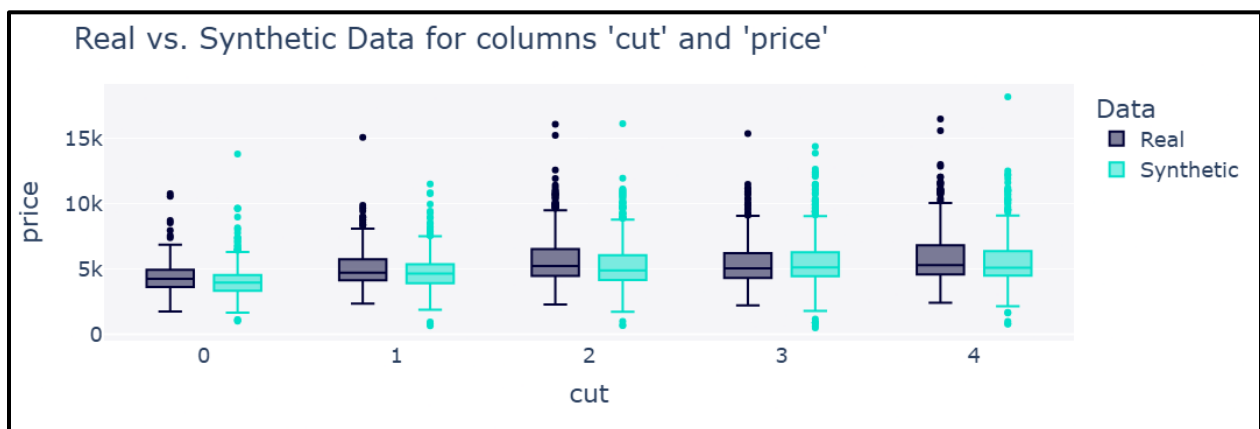


Figure 26: 'cut' vs. 'price' where 'carat' ≈ 1 for Real and Synthetic Data

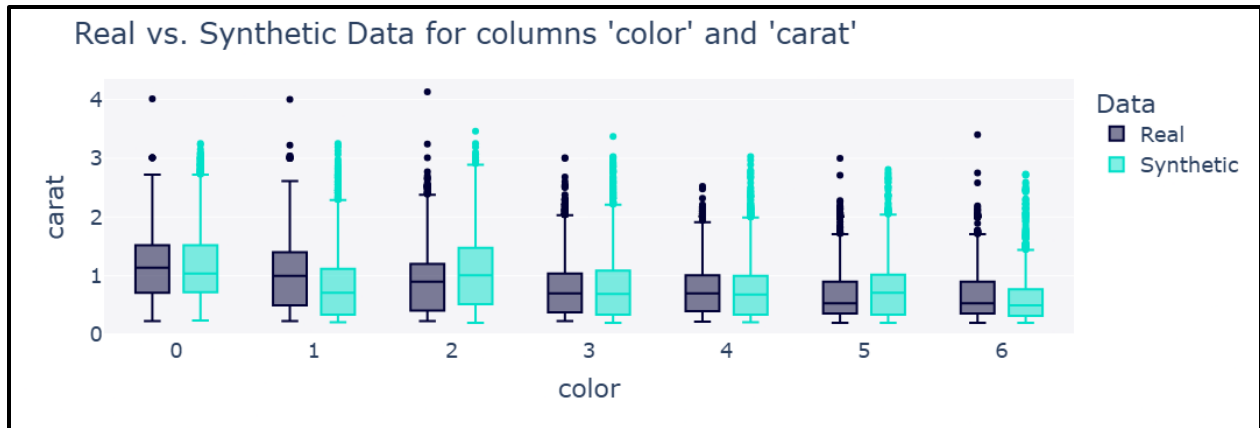


Figure 27: 'color' vs 'carat' for Real and Synthetic Data

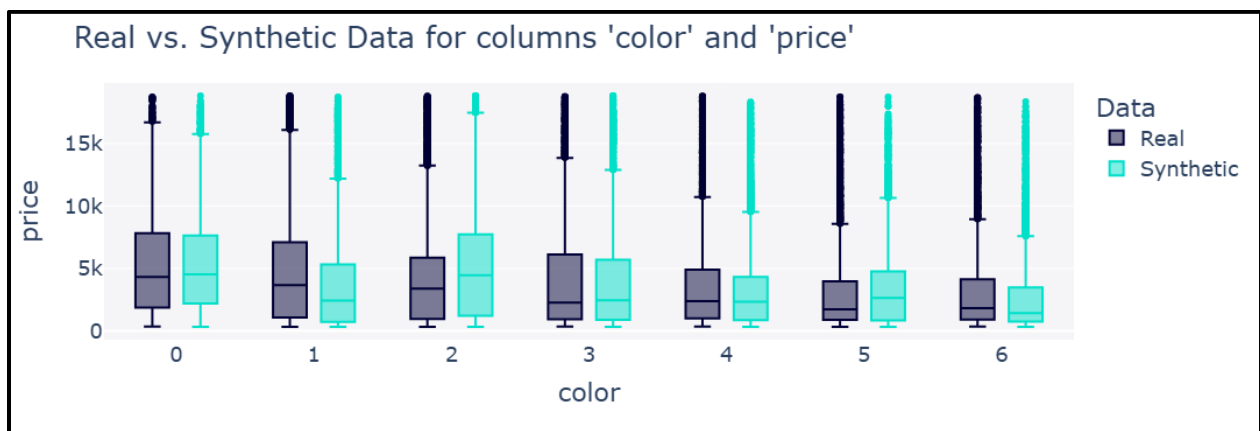


Figure 28: 'color' vs 'price' for Real and Synthetic Data



Figure 29: 'color' vs 'price' where 'carat' ≈ 1 for Real and Synthetic Data