# 1 Introduction

## 1.1 Abstract

Imputation is a process of replacing missing data with substituted values. Missing data or imputation data with not substituted values can create problem with analyzing data which may affect the representativeness of the results. our goal is to create a tool that find the best imputation to those missing values. In our solution we tried a different imputation methods and test them with statistical tools such as dataset distribution, ks test, kurtosis and skewness. According to those tools we chose the best imputation. we tried this methods on 4 different dataset and the result was that knn is the best solution for imputation data.

## 1.2 Problem description

Data science is the basis for everything in today's world. With huge amounts of information flowing at any given moment, the significant task is to be able to mine the relevant information. Missing data is one of the most common problems in data processing. this problem is part of the cleaning data step of the Data Science Pipeline. At the cleaning data step we solving problems that may inflect our result such as duplicate parameters, missing values, or irrelevant information. The intuitive solution is to the missing data is to delete them (Azar, 2002), but this leads to two problems - on a practical level, we have to deal with situations where most of the data has missing information, and deleting it will leave us with limited data. At the statistical level, there may be certain characteristics of the data that cause it to be missing. Variables can be completely randomly missing - MCAR, missing randomly but with a depends on available information - MAR, and missing not randomly - MNAR. deletion missing value will lead to ignoring these characteristics.(Little  Rubin, 2002).

The first question when we come to address missing data is how to deal with it. There are many methodologies that try to deal with missing data (SchaferGraham, 2002), each with its own pros and cons. The second question is how to evaluate the differences between the different methods and choose the one that fits our data. In this article we will try to review a number of different methods and different ways of evaluation and compare them.

## 1.3 Related Work

Imputation data task is a known problem and there is a lot of information and technique to deal with this problem, In opposite to those techniques that choosing the method by plotting the data and then analyse it, we are trying to analyse the data automatically and return the best method based on the dataset. What we're innovating is the way dealing with the data after the imputation. we using skewness, kurtosis, ks test and sum square error to get the best method.

# References

[1] Azar, B. (2002). Finding a solution for missing data. Monitor on Psychology, 33(7), 70. Little, R. J.,  Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley  Sons.

[2] Schafer, J. L.,  Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147.

[3] Missing Data Imputation for Geolocation-based Price Prediction Using KNN–MCF Method Karshiev Sanjar, Olimov Bekhzod, Jaesoo Kim, Anand Paul and Jeonghong Kim *

# 2 Solution

## 2.1 General approach

Given a dataset with missing value, our solution is a system that take diffrenet imputing methods and chose the best method correspondingly to the dataset distribution. given a dataset we taking put 5% of each feature, imputing the data with all the methods above and with dataset distribution, ks test, kurtosis and skewness find the best way to impute.

In this section we will describe different methods for treating missing values in dataset.

.**Common methods for completing missing data**

## 2.2 statistical methods

**Mean**. The most popular method is mean - replacing the missing value with the average value of this variable. The problem with this method is that it assumes that there is no dependency between this variable and other variables in the data, and can impair finding an existing correlation.
$\bar{x_{ij}} = \sum_{i:x_ij \in C_k} \frac{x_{ij}}{n_k} 1$

**Median**. Similar option is the median - replacing the missing value with the variable median. Here too the problem is similar to the mean method, and in fact we fill in all the missing spaces with the same value and ignore its probability of changing according to its other characteristics.
$\hat{x_{ij}} = median_{i:x_{ij} \in c_k} x_{ij} 2$

**Most Frequent**. another option to imputing values in the dataset is impute the most frequent value at the missing places. As the mean method this method also have problem with the correlations between features. this method works with categorical features, but the result with numerical features aren't good enough.

**Zero Values/Constant Values**. In this method the missing data replace with zero values or constant (some value Predetermined). For convenience of this assignment we decided to try this method with zero values. Also in this method there isn't correlation between features.

**Regression**. Another option is to use regression. Based on the analysis of the relationship between the various variables in the data, we will create a complex regression equation based on the relationship between each of the variables and the column of missing variables, and based on the existing values we will predict the missing variable. This method deals with the problem of assuming a lack of connection that we presented in the previous sections - but raises the opposite problem because completion by regression is based on assuming the existence of a connection between the variables, when in practice it does not necessarily exist, or exists but with weak intensity.

## 2.3 Machine Learning methods

**Knn algorithm**. k-nearest neighbors algorithm (k-NN) is a ML algorithm for classification. given an input of new example the algorithm find the most common class to this example among k neighbor. we use this algorithm to predict the missing data among k's closet neighbours from the existing data. As opposed to the previous methods, this method based on feature similarity which save the correlations between features.

after the imputation, we want to automatically chose the best methods, due to that we use sum square error, KS test and kurtosis and skewness to analyse the data. we choose the method that those tests return as the best.

## 2.4 Sum Of Squares Error

Probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. each feature have a distribution, and we like to preserve our feature distribution. Due to that we check the original data distribution and compare it to each of the feature distribution after imputing by the different method and return the methods that preserve the distribution. there are few way to check the distribution one of them is KS test as will describe next subsection.
we check the distrubtns with sum of squares error. The error is the difference between the observed value and the predicted value.

### 2.5 KS test

the Kolmogorov–Smirnov test (KS test) is a nonparametric test of the equality of continuous probability distributions that can be used to compare a sample with a reference probability distribution. we used KS test to check the original dataset distribution and the new dataset distribution (the dataset after imputing) and return the method with the maximum p-value. the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. which means that if the p-value is higher the probability of the new distribution to be same as the original is higher.

### 2.6 kurtosis and skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. we want to find which method give us the minimum changes in the kurtosis and skewness, we compare the original dataset to the dataset after imputing and return the method with the lowest changes.

we use those four - distribution, ks test and kurtosis and skewness and chose the method that on average bring the most similar result.

## 3 experimental results

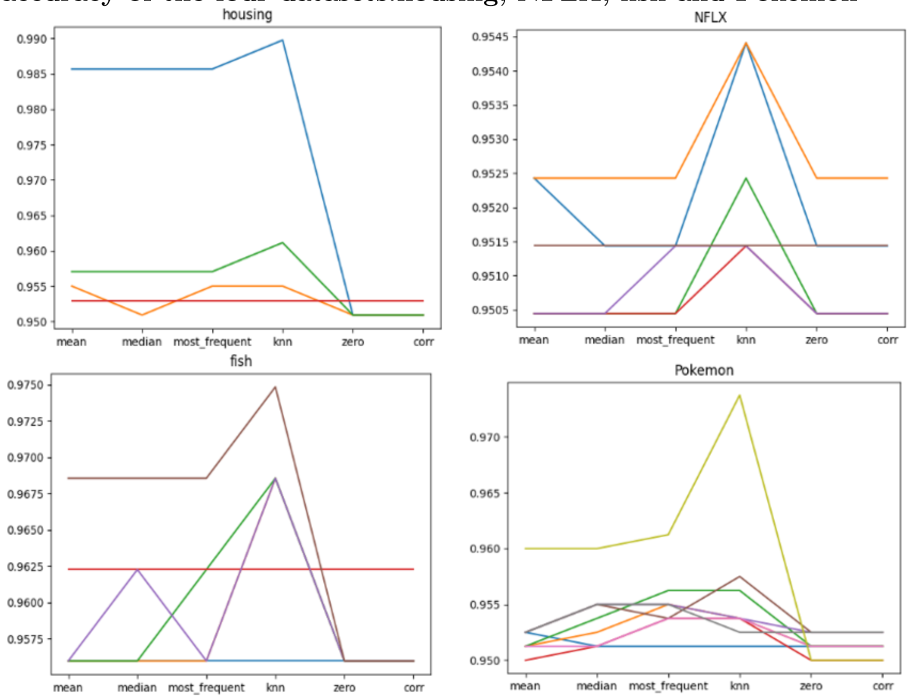
to chose the best method we count each method number of best solution we received. we can see at the table below knn had the higher amount of counting. Due to that knn return as the best method.

**dataset 'housing' - frequent best result for each one of 10 features.**

| Country List | | | | |
|---|---|---|---|---|
| method | Sum Square Error | KS test | Skewness | Kurtosis |
| Mean | 0 | 0 | 2 | 0 |
| Median | 0 | 0 | 1 | 2 |
| Most Frequent | 1 | 0 | 1 | 0 |
| Zero | 1 | 0 | 0 | 0 |
| Regression | 0 | 0 | 0 | 0 |
| KNN | 6 | 9 | 5 | 7 |

To check to correctness of the chosen method we run test on four datasets: housing, NFLX, fish and Pokemon. we check the system by checking the accuracy. accuracy is how close or far off a given set of measurements are to their true value. high accuracy mean we close to the true value, it mean we have high skew. we want our chosen method to have the higher accuracy.

**accuracy of the four datasets:housing, NFLX, fish and Pokemon**



At the figure above we can see that most of the features have the highest accuracy when using imputation of knn.

## 4 Discussion

Imputation of data is important task in the DS pipeline, it effect our data analyses and have to be treated correctly. We proposed techniques to deal with missing data in this paper, statistical and ML methods. the result that we get as the best method is the KNN algorithm, we tested the accuracy of each of the method and get that the system is working and knn is the best method for each one of the four different datasets that has been tested. our approach make it easier to chose the best method and using differnet staticial methods to find the best imputation. Despite obtaining good accuracy for the prediction, we believe that various improvements can be made in the future, such as the selection of the most important features can be performed using deep learning.