Distributions Similarities

Tabular Data Science Final Course Project

Simon Raviv (ID. 312847478), Adam Gavriely (ID. 309677284)

Submitted as final project report for TDS course, BIU, 2022

Abstract

A key part of the data science pipeline is understanding how the data is distributed, and this can be used for a variety of purposes. Statistical hypothesis testing is a popular method of examining data distributions. Despite their wide use, these statistical tests have poor performance in various situations. For example, comparing samples with the same distribution but slightly different parameters as well as performing tests on similar but not identical distributions. Furthermore, the stability of the test is affected by various factors. p-values are affected by these problems significantly. Our method is to compute normalized Root Mean Squared Error (RMSE) between the KDEs of both distributions in order to estimate distribution similarity between them. Our experimental results in this paper demonstrate that our method can handle the problems with statistical tests and can replace or serve as a complementary tool for comparing distributions. For multiple samples of the same generator, our method shows stable and robust results. As opposed to statistical hypothesis testing, our method have little to no influence from random factors.

1 Problem Description

In data science projects, statistical tests are frequently used for null hypothesis testing and to determine whether data comes from a specific population or which distributions are identical [lecture 4]. When our data is not distributed normally, which is often the case, we must use a non-parametric test like two sample Kolmogorov–Smirnov test [lecture 8]. These non-parametric statistical tests perform poorly in several use cases despite their widespread use. It is possible to get significantly small p-values when two samples are drawn from the same distribution using slightly different parameters as seen in figure 1 where we used 2 normal distributions that differ only by their std (1 and 1.05 accordingly), or from two distributions that are quite similar like in figure 2. In addition, by changing the sample size, the statistical test, or resetting the random seed, p-value results may differ significantly. Additionally, those tests are unable to detect similarities among different distributions. And as seen in figure 3, testing several samples from the same generator can produce very different p-values, sometimes even below the minimum acceptable level.

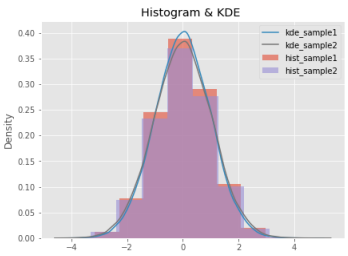


Figure 1: Same distribution, slightly different parameters

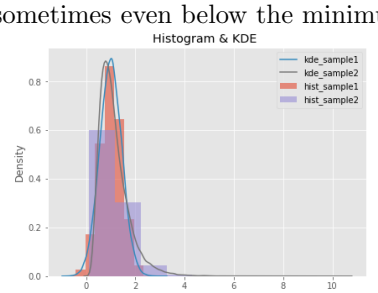


Figure 2: Similar looking distributions

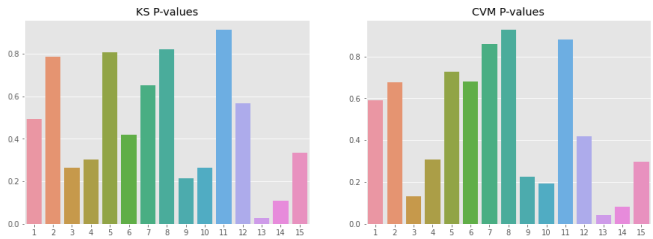


Figure 3: Same distributions,unstable p-values

2 Solution Overview

Our first step involved fitting real-world data to theoretical distributions with Maximum Likelihood Estimation (MLE) looking for the top theoretical distributions. The Fisher method to combine p-values is used to score the fittings based

on p-values from KS and Cramer-Von Mises (CVM) statistical tests. We found that not every data set can be fitted to a theoretical distribution. While not all samples can be fitted to some theoretical distributions, those that can are still not producing very high p-values (shown in 3.2).

We chose to focus on similarities between empirical distributions of continuous data. Using two-sample tests [lecture 8] we realized that p-values are very low, even when the distributions are quite similar.

Our goal was to develop a stable and robust measure that lets users compare similarities between distributions. The basic concept of the proposed method is to use the Kernel Density Estimation (KDE) functions [lecture 7] of both distributions and calculate the normalized Root Mean Squared Error (RMSE) between their values. Furthermore, we added the ability to scale the data using min-max normalization, so that we could compare distributions that comes from different domains (e.g. height and weight).

Algorithm 1 The Raviv-Gavriely(RG) method for comparing distributions similarity

Input $sample1, sample2$: 2 data samples
$scaling$: boolean to determine if scaling is necessary
Output rg_score : score in range [0,1], the percentage of similarity between distributions

```
if scaling is True then
    sample1 ← scaled(sample1)                    {min-max scale to range [0,1]}
    sample2 ← scaled(sample2)
end if
kde1 ← KDE(sample1)                              {Calculate the KDE function for each sample}
kde2 ← KDE(sample2)
data ← sort(samples1 ∪ samples2)                 {Sorted container of all the data points}
values1 ← kde1(data)                             {Evaluate the values of the KDEs on each point}
values2 ← kde2(data)
rmse_score ← RMSE(values1, values2)              {Compute RMSE between the KDEs values}
rg_score ← normalize(rmse_score)                 {Normalize the result to range [0,1]}
return rg_score
```

3 Experimental Evaluation

3.1 Intro

We conducted experiments to demonstrate the current problem and how our solution can resolve it. Fitting real-world data to theoretical distributions was our first step. Then we used synthetic data to show different aspects of the problem and our solution, and then we tested it on real-world data. To prevent any random effects between tests and to maintain consistent results, we handled the random seed separately in each experiment. In this report, we present just a subset of our experiments. To view the full experimental results, please refer to the notebook.

3.2 Top-K Theoretical Distributions

First we experimented with fitting real-world data to theoretical distributions and looked for top k theoretical distributions. It appears that not all datasets fit to a theoretical distribution (figure 5), and even those that do are not producing a very high p-values as seen in figure 4.

As mentioned above, this is a complementary perspective, not the main focus of the project.

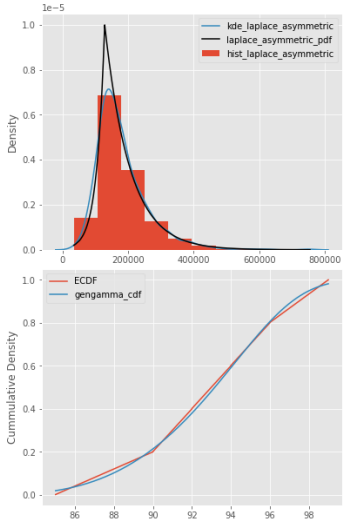


Figure 4: Laplace asymmetric distribution fitted to house prices

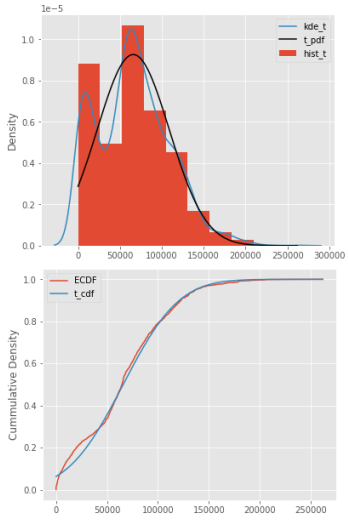


Figure 5: T distribution fitted to base salary pay in California

Distribution Name	P-Value
Laplace asymmetric	0.6406
T	$4.7430e-09$

3.3 Synthetic Data

We synthesized different scenarios to illustrate how our method compares to KS and CVM statistical tests for synthetic data samples.

3.3.1 Experiment 1 - Same generator with positively skewed distribution

This experiment shows the comparison between 2 samples of size 1,000 from the same Gamma distribution generator to see how the tests would behave around positively skewed distribution [lectures 5-6]. Considering the two distributions drawn from the same generator and are indeed very similar (figure 6) we would expect a much higher p-value. Our method scored the highest with 95% similarity.

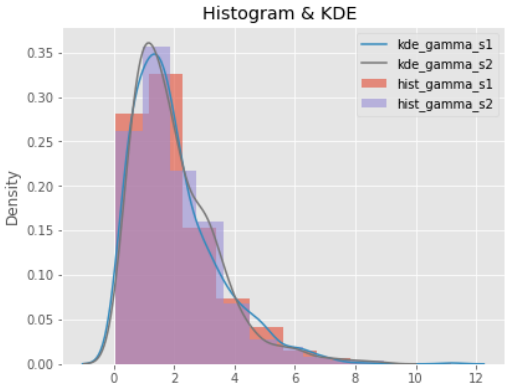
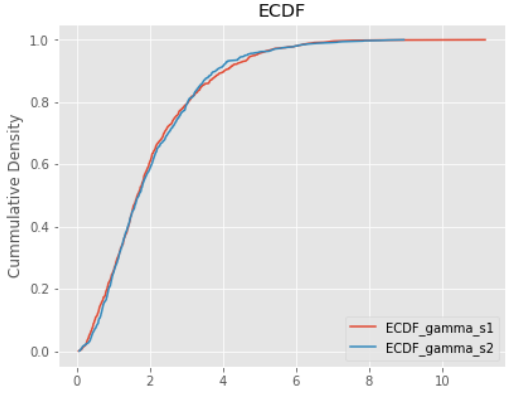


Figure 6: Histogram and KDE



Metric	Value
RG Similarity	95.4234%
KS P-Value	0.432608
CVM P-Value	0.464980

Figure 7: ECDF

3.3.2 Experiment 2 - Normal distribution similarity with data scaling

In this experiment, we compare 2 samples of 1,000 from the same distribution function, but with different parameters, i.e. different domains of values. There are two distributions $X \sim \mathcal{N}(170, 10^2)$ and $Y \sim \mathcal{N}(80, 5^2)$. In figure 8, we demonstrate how scaling the data to the same domain affects the similarity. We first compared without scaling and found relatively small results, but not as small as the p-values, which are close to 0. Then we compared the same distributions but scaled the samples to the same domain and saw that our method score was up to 77.7% similarity.

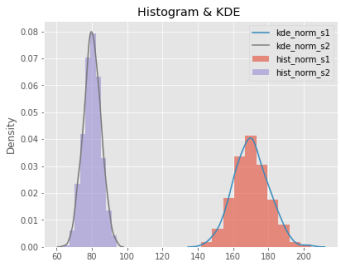


Figure 8: Histogram and KDE before and after scaling

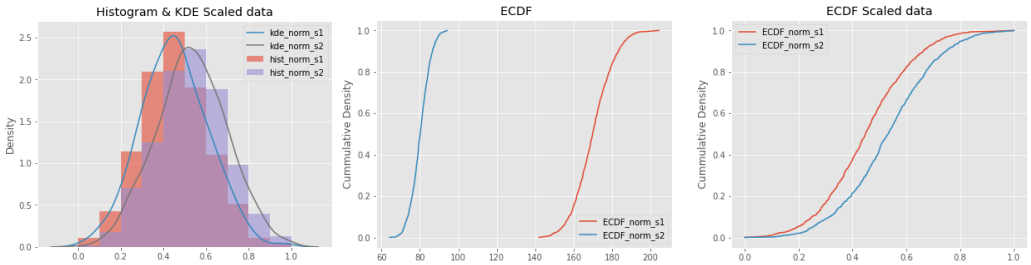
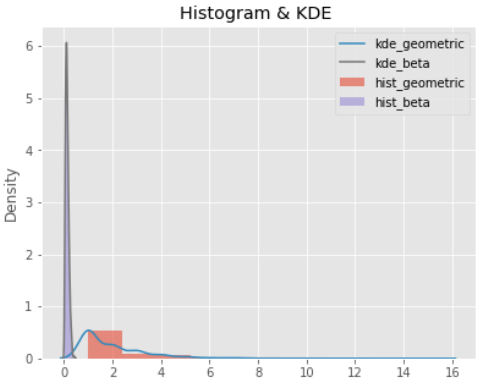


Figure 9: ECDFs before and after scaling

Metric	Value
RG Similarity	40.5981%
RG Similarity (with scaling)	77.7008%
KS P-Value	0
CVM P-Value	$4.03740e-08$

3.3.3 Experiment 3 - Different generators

This experiment shows the comparison between 2 samples of size 1,000 from the different distribution generator. The first is from the Geometric distribution and the second from the Beta distribution. We see that both p-values fail to detect even the slightest similarity, our method scored 24.4% which shows that even though they are not the same distributions there is a small similarity between them as seen in figure 10.



Metric	Value
RG Similarity	24.4889%
KS P-Value	0
CVM P-Value	$6.55374e-08$

Figure 10: Histogram and KDE

3.3.4 Experiment 4 - Stability on large sample size

In this experiment we sampled twice from the same Normal distribution generator with the sample size of 1,000 and compared the results. We did the same thing 10 times and compared the results between every iteration. We see that even with larger sample sizes the p-values still change drastically between iterations as shown in figure 11. CVM test lowest score was even below the widely used acceptance level of 0.05. Again our method was the most consistent and stable as shown in table 1.

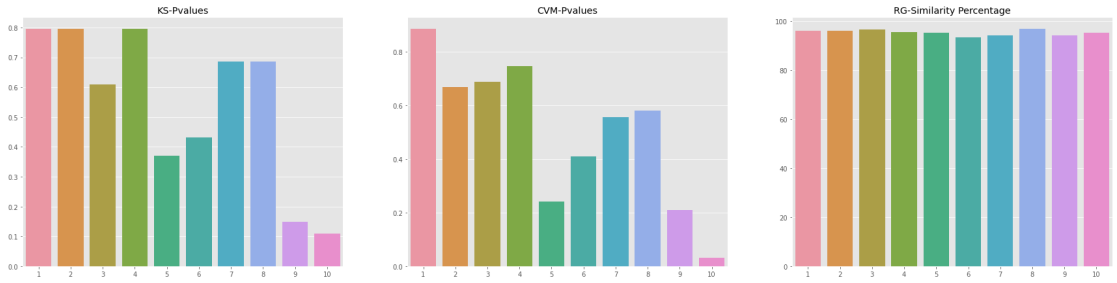


Figure 11: KS P-values, CVM P-values, RG Similarity Percentage

Method	Min Value	Max Value	Mean Value	Standard Deviation
KS P-Value	0.10829	0.79466	0.54245	0.24938
CVM P-Value	0.03170	0.88441	0.50117	0.25698
RG Similarity	93.28%	96.70%	95.26%	1.05%

Table 1: Stability results statistics on same generator with 1000 sample size

3.3.5 Experiment 5 - Stability on gradual change in parameters

In this experiment we started with a distribution from $X \sim \mathcal{N}(0,\, 1)$ and sample size 500. Then we gradually increased the mean and std of the distribution by 1 with each iteration and compared to the original distribution. We showed that when the distribution gradually change so does our score (figure 12. Because of the low scores of the statistical tests we used $f(x) = \frac{-1}{log(x)}$ to demonstrate the gradual changes. CVM showed non gradual changes.

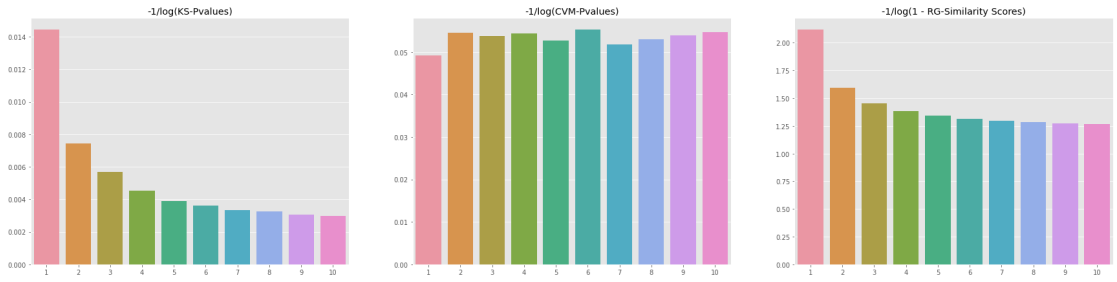


Figure 12: KS P-values, CVM P-values, RG Scores

3.4 Real Data

We have used 4 real world dataset to demonstrate our metric. This section of the experiments shows how real world datasets compared with our method vs KS and CVM statistical tests. For further information, follow the links below.

- Dataset 1 - Houses price prediction
- Dataset 2 - Cancer detection
- Dataset 3 - Stress detection in sleep
- Dataset 4 - Salaries in san fransisco

3.4.1 Experiment 1 - Dataset train/test split use case

This experiment demonstrates distribution similarity between two groups sampled from the same data. This is useful when splitting the dataset to train 80[%], test 20[%] sets, in this scenario its important to verify the sets distributes the same for model training. Figures 13 and 14 refers to feature *SalePrice* - The property's sale price in dollars, from Houses

price prediction, where the dataset sample size is 1,460 items. Figures 15 and 16 refers to feature *BasePay* - The base pay of the salary, from Salaries in San Francisco, where the dataset sample size is 10,000 items.

We can see our metric gives high similarity between the samples, while the p-values in the statistical test are above the common confidence level of 0.05, but yet they don't give high p-value.

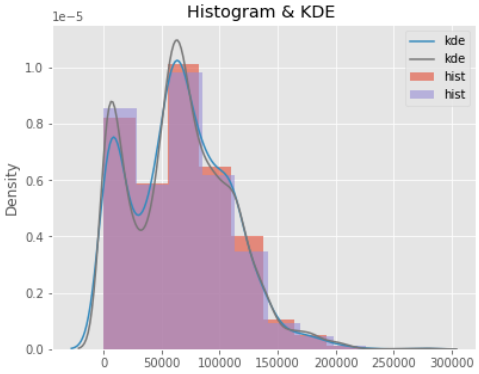
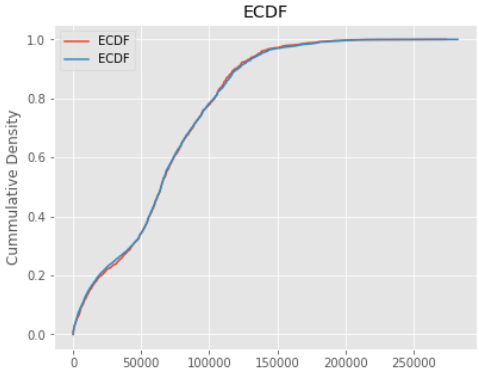


Figure 13: Histogram and KDE



Metric	Value
RG Similarity	95.9095%
KS P-Value	0.77795
CVM P-Value	0.69914

Figure 14: Two ECDFs

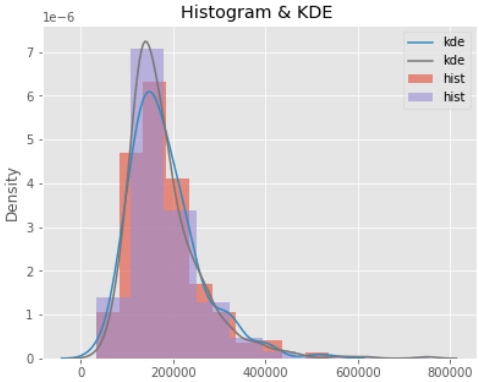
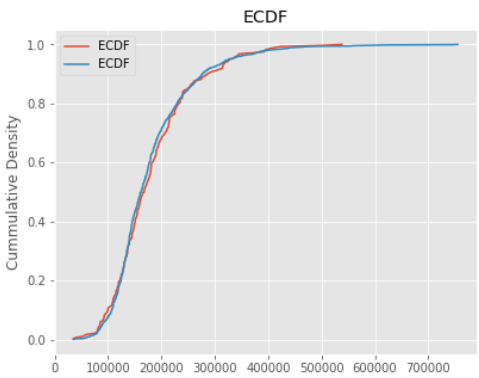


Figure 15: Histogram and KDE



Metric	Value
RG Similarity	94.4035%
KS P-Value	0.78039
CVM P-Value	0.8453

Figure 16: Two ECDFs

3.4.2 Experiment 2 - Model training evaluation K-Fold cross validation use case

This experiment demonstrates distribution similarity between 10 random samples from the same data. Each sample split to train 80[%], test 20[%] sets. This is useful when doing K-Fold cross validation, in this scenario its important to verify the different folds have train/dev split as close as possible data distributes for accurate model evaluation over the folds. The feature is *SalePrice* - The property's sale price in dollars, from Houses price prediction. The dataset sample size is 1400 items.

We see that even though its different random samples from the same dataset, the p-values changes drastically between iterations as shown in figure 17. KS and CVM tests lowest scores were even below the widely used acceptance level of 0.05. Again our method was the most consistent and stable as shown in table 2.

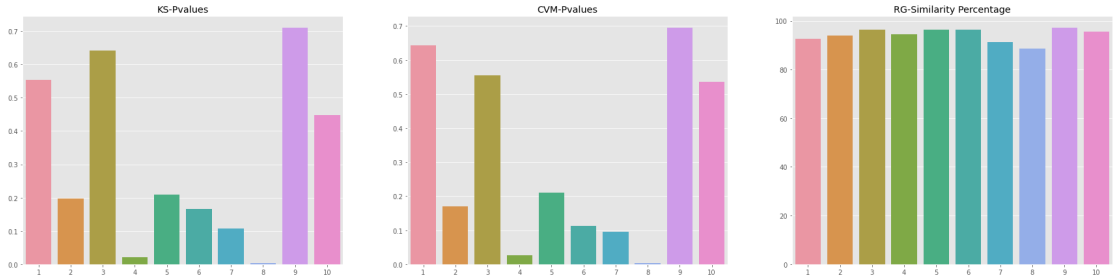


Figure 17: KS P-values, CVM P-values, RG Similarity Percentage

Method	Min Value	Max Value	Mean Value	Standard Deviation
RG Similarity	88.74%	97.08%	94.29%	2.54%
KS P-Value	0.00412	0.70924	0.30569	0.24669
CVM P-Value	0.00407	0.69422	0.30481	0.25656

Table 2: Stability results statistics

3.4.3 Experiment 3 - Comparing different features use case

This experiment demonstrates distribution similarity between two features in the same dataset (can be any two data samples). Figures 18 and 19 refers to features *sr vs t* - Snoring range vs body temperature, from Stress detection in sleep, where the dataset sample size is 569 items. We see low similarity percentage for all the methods, even when scaling. This is reasonable, as we can see in the figures they distribute very differently.

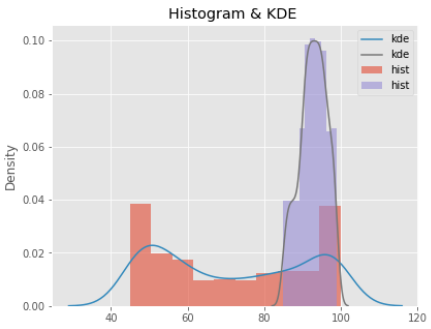


Figure 18: Histogram and KDE

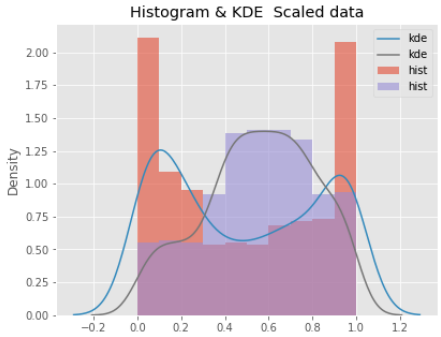


Figure 19: Histogram and KDE Scaled

Metric	Value
RG Similarity	49.6575%
RG Similarity (with scaling)	47.8876%
KS P-Value	$4.8178e-133$
CVM P-Value	$4.03740e-08$

4 Related Work

There are several existing techniques that can be used to understand data distribution similarities, we will discuss two of them.

4.1 Statistical Tests

Statistical tests can be used as distribution tests to identify the probability distribution that the data follow. Distribution tests are hypothesis tests that determine whether your sample data were drawn from a population that follows a hypothesized probability distribution. Like any statistical hypothesis test, distribution tests have a null hypothesis and an alternative hypothesis. For distribution tests, the smaller the p-values the higher the chance you can reject the null hypothesis and conclude that your data were not drawn from a population with the specified distribution [lecture 4]. However, we want to identify the probability distribution that our data follow rather than the distributions they don't follow. Consequently, distribution tests are a rare case where you look for high p-values to identify candidate distributions. Therefore, they don't really give distributions similarity level.

4.2 Kullback–Leibler Divergence (Relative Entropy)

The Kullback–Leibler divergence, $D_{KL}(P \parallel Q)$, is a statistical distance: a measure of how one probability distribution Q is different from a second, reference probability distribution P. A simple interpretation of the divergence of P from Q is the expected excess surprise from using Q as a model when the actual distribution is P. It is a distance, but it is

asymmetric in the two distributions. In the simple case, a relative entropy of 0 indicates that the two distributions in question have identical quantities of information. In addition, it gives a distance that is 0 or grater, while our method gives a score between [0,1] for similarity.

5 Conclusion

5.1 Findings

In conclusion, we found that p-values from statistical tests are not very stable and can not be used as a metric to compare similarities between distributions. Many factors can change the p-values drastically, for example the sample size, type of distribution and the sample domain. Our method proved to be a more stable method to compare distribution similarities and from the empirical results we saw that it performs at least as good as a statistical test and very often much better when we want to find how similar one distribution is to the other. In addition. we saw that when using synthetic data it is much easier to control the experiment and the results are much more nicer and clearer. In the real-world the data does not always distributed as nicely.

5.2 Project Insights

5.2.1 Technical Insights

The project covered several topics in detail, and some of them are also covered in the course lectures, mainly in the Exploratory Data Analysis and Statistical Correlation Measures lectures. We started with a simple metric to get the theoretical distributions the data most likely drawn from. During the work on this metric we learnt about fitting of parameters of different theoretical distributions to the data using MLE. In addition, we learnt that it is possible to use p-values from multiple tests and combine them together, specifically we used the Fisher method. During this process, we got familiar with multiple types of distributions.

We got deeper understanding about non parametric statistical tests and p-values, what p-values are and also what they are not, it was not an easy concept to grasp. As statistical tests are not really a metric for similarity of distributions and suffer from several issues discussed above, we then thought how to leverage KDE of data to understand its distribution and use it as the base for our method for distribution similarities. This got us thinking of the new method in order to fulfill the goal of the project.

5.2.2 Other Insights

- **Data alone is not enough.** Data alone almost never gives value, it takes work to explore the data, understand it and extract the value from the data.
- **There is no silver bullet.** There are a lot of methods to handle and study the data, and many statistical tests that can be done. There is no "magic" test that fits every case.
- **Always question your results.** No matter what statistical tests or methods you use, never take the results as granted. Always try to explain to yourself why you got this results and see if it makes sense to you.