4. Inferential Statistics:

Following up from the data wrangling step, we will explore the dataset from a statistical point of view. As a review, the dataset is comprised of 5435 sound files indexed by a numeric ID that we want to classify into ten classes or types of urban sounds. Ground truth labels are stored in a dataframe called *labels*. To start, we are interested to know how many total files are in each class. After grouping and counting ids, we get the following table (Table 1) of number of files present in the dataset for each of the ten classes:

Table 1: Number of files in each class

Class	**Number of file samples**
air_conditioner	600
car_horn	306
dog_bark	600
children_playing	600
drilling	600
engine_idling	624
gun_shot	230
jackhammer	668
siren	607
street_music	600

We will conduct our statistical analysis through two separate fields of view. First, we will treat the sample space as consisting of the different sound files. In this view, all samples are independent, and we can apply classical statistical methods in our analysis. The second view we implement is the time series modeling approach where we model our sound amplitude or frequency data as observations from a stochastic process.

What follows is a description of both views in two separate sections.

4.1. Part 1: Independent Samples:

Looking at the data storytelling plots from before, we come up with the following hypothesis: the means of “jackhammer”, “air_conditioner” and “street_music” wave plot (amplitude vs time) sample distribution is the same. In other words, we can state the null hypothesis as the difference in sample distribution means of the three classes is zero.

To test this claim, we will reduce the varying amplitudes over time to the mean amplitude for each file sample then apply the t-test on the somewhat independent mean samples of each class. We run this t-test three times to compare the means of the three class distributions.

Claim 1: The difference in sample distribution means of “jackhammer”, “air_conditioner” and “street_music” classes is zero.

Tests and Results:

1. Test: ttest_ind(jackhammer_mean_files, air_conditioner_mean_files)

   Result: t-statistic = 2.8805317026359107, p-value = 0.004036935822494086

Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis of equal distribution means.

2. Test: ttest_ind(jackhammer_mean_files, street_music_mean_files)

   Result: t-statistic = -1.013447040166551, p-value = 0.31104032436124646

Since the resultant p-value is greater than 0.05 (5%), we can neither reject nor accept the equal means hypothesis. The two groups might have equal means.

3. Test: ttest_ind(street_music_mean_files, air_conditioner_mean_files)

   Result: t-statistic = 2.5641908015042962, p-value=0.010462734082099105

Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis of equal distribution means.

4.2. Part 2: Time Series Approach:

We will answer the following questions in this section:

1. Investigate the stationarity of “street_music” and “siren” using Augmented Dickey Fuller Test. (Null Hypothesis is that data are non-stationary)
2. Are “jackhammer” and “air_conditioner” more stationary than “street_music”?
3. How similar (or correlated) are the wave plots of “drilling”, “engine_idling” and “car_horn”?

A dataset is stationary if the mean is constant over all time points. If a time series has a unit root, it shows a systematic pattern that is unpredictable. The Augmented Dickey-Fuller Test (James G. MacKinnon, 2010 "Critical Values for Cointegration Tests, Economics Department, Queen's University) is used to test for a possible unit root or the stationarity of a time process. The critical values for the test are as follows for 1%, 5% and 10% confidence levels.

```
{'1%': -3.4305546957800765,

 '5%': -2.86163047020448,

 '10%': -2.5668181544937676}
```

Before we run any statistical tests, we generate a representative wave plot sample for each class distribution by using the mean amplitude of all samples at a time point as the amplitude value of the class at that same time point. Since not all file samples have the same time duration, we do the averaging for the smallest time period present in the class sample files. Table 2 shows the final number of data points computed for each class.

Table 2: Number of data values in the final sample for each class

Class	**Number of data values in the representative sample**
air_conditioner	97920
car_horn	2205
dog_bark	5284
children_playing	55125
drilling	18383
engine_idling	33805
gun_shot	7333
jackhammer	18720
siren	26001
street_music	32000

Next, we make the following claims:

Claim 2: There is a high degree of stationarity in the wave plots of "street_music" and "siren".

Tests and Results:

1. Test: adfuller(street_music_rep_sample)

   Result: Statistic = -21.61937687757208, p-value = 0.0

Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis that the "street_music" data are non-stationary and there is no trend in the time series.

2. Test: adfuller(siren_rep_sample)

   Result: Statistic = -18.827392026204237, p-value = 2.022129131934057e-30

Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis that the "siren" data are non-stationary and there is no trend in the time series.

Claim 3: The wave plots of "jackhammer" and "air_conditioner" are more stationary than "street_music".

Tests and Results:

1. Test: adfuller(jackhammer_rep_sample)

   Result: Statistic = -13.869743769856765, p-value = 6.497786183782305e-26

Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis that the "jackhammer" data are non-stationary and there is no trend in the time series.

2. Test: adfuller(air_conditioner_rep_sample)
   Result: Statistic = -18.912595716668758, p-value = 0.0

Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis that the "air_conditioner" data are non-stationary and there is no trend in the time series.

3. Test: adfuller(street_music_rep_sample)
   Result: Statistic = -21.61937687757208, p-value = 0.0

Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis that the "street_music" data are non-stationary and there is no trend in the time series.

Comparing the stationarity of the three classes, "street_music" has the largest negative statistic from the critical value which implies that "street_music" is more stationary than "jackhammer" and "air_conditioner".

Claim 4: The wave plot of "engine_idling" is the most stationary compared to "drilling" and "car_horn".

Tests and Results:

1. Test: adfuller(engine_idling_rep_sample)

   Result: Statistic = -12.348556688899007, p-value = 5.896139209666792e-23

   Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis that the "engine_idling" data are non-stationary and there is no trend in the time series.

2. Test: adfuller(drilling_rep_sample)

   Result: Statistic = -17.573260000963817, p-value = 4.0415103709846376e-30

   Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis that the "drilling" data are non-stationary and there is no trend in the time series.

3. Test: adfuller(car_horn_rep_sample)

   Result: Statistic = -6.937221617019703, p-value = 1.0471009403741136e-09

   Since the resultant p-value is less than 0.05 (5%), we can reject the null hypothesis that the "car_horn" data are non-stationary and there is no trend in the time series.

Comparing the stationarity of the three classes, "drilling" has the largest negative statistic from the critical value which implies that "drilling" is more stationary than "engine_idling" and "car_horn".

We are interested next in finding the correlation between the classes. To test this, we compute the Pearson Correlation Coefficient between the representative samples of the classes in interest. We make the following claim:

Claim 5: The wave plots of "engine_idling" and "drilling" are highly correlated.

Test and Result:

1. Test: stats.pearsonr(engine_idling_rep_sample[:com_len], drilling_rep_sample[:com_len])

   Result: Correlation coefficient = -0.0012251173053355313, p-value = 0.8680822072537137

   The correlation results indicate that the "engine_idling" and "drilling" classes are not correlated. Since, the p-value is higher than 0.05 (5%), we cannot reject the null hypothesis that there is no correlation between the two sample classes.

Following our data exploration work, we are interested to examine the linear relationship between the wave plots of two sets of classes: "dog_bark", "gun_shot" and "children_playing", "street_music". We end our analysis by investigating that relationship using the Pearson Correlation measure as before.

Question 1: What is the correlation between the wave plots of "dog_bark" and "gun_shot"?

Test and Result:

Correlation coefficient = 0.05881049379426529, p-value = 1.8884083840078154e-05

Question 2: What is the correlation between the wave plots of "children_playing" and "street_music"?

Test and Result:

Correlation coefficient = -0.0009760590800338108, p-value = 0.8613971097374443

The results above indicate no correlation between the classes in both sets. The first result appears to be more significant than the second due to the lower p-value of the correlation test.

4.3. Conclusion:

In this section, we applied inferential statistics techniques to our dataset after the data exploration stage. We applied various statistical tests in two viewpoints with different assumptions to wave plot data i.e. amplitude values versus time. It would also be interesting to apply similar inference techniques to 2D spectrogram data i.e. amplitude values versus frequencies and time. However, this requires some independent component analysis of the spectrogram matrix to find the independent frequency and time components. Non-negative Matrix Factorization is a popular technique that can be used for this goal.