MATH 6333 CASE STUDY 1

AKUA SEKYIWAA OSEI-NKWANTABISA
Student ID: 20674356

2024-03-02

1 Introduction

Tinnitus is among the most distressing hearing-related conditions. This could have a negative impact on people's quality of life and be a severe problem. The UK has created an internet cognitive behavioral therapy intervention in order to enhance access to research-based tinnitus treatment (referred to as the treatment here in the data set). The Tinnitus Functional Index (TFI score) is the primary assessment instrument for determining the degree of tinnitus suffering before and after therapy. Investigating the genesis of tinnitus in light of the following discoveries is one of the primary objectives of this study, along with gaining practical experience using supervised learning techniques on this set of real-world data.

2 Literature Review

This review research includes pertinent theories and writings on tinnitus. The study shows that tinnitus research has been pivotal for a long time and is a challenging task. Researchers indicate that there are not enough trials to definitively identify whether most tinnitus interventions recommended in clinical practice are effective. Tinnitus is one of the most common otologic diseases with a condition in which an individual recognizes sounds in the absence of external sound stimulation[1]. There are several subtypes of tinnitus such as conductive tinnitus, sensorineural hearing loss, and vascular tinnitus. Conductive tinnitus can occur because of middle ear origins such as ear infections, tympanic membrane and ossicular chain problems, glomic tumors, myoclonus, and tonic tensor tympani syndrome. Sensorineural tinnitus is accompanied by sensorineural hearing loss, which is the most common type of tinnitus. It can be associated with presbycusis, metabolic problems such as diabetes mellitus, hypothyroidism, dyslipidemia, anemia, vitamin and mineral deficiencies, and noise exposure. Vascular tinnitus can be produced by the turbulence of blood flow transmitted to the cochlea [2 3]. The most frequent causes of tinnitus associated with hearing loss are noise-induced hearing loss and presbycusis [4]. The mechanism of tinnitus can be explained based on nerve-fiber activity in the temporal lobe of the cortex—the same as the perception of all sound. The activity can be caused by several mechanisms. The spontaneous activity of neurons in the auditory system included deafferentation and central changes as well as

an increase in cross-fiber correlation. The case of tinnitus due to hearing loss can be explained by the cochlea origin, but the central origin cannot be ignored [5] [6]. According to the diverse mechanisms of tinnitus, previous studies emphasized the importance of subgrouping tinnitus patients for treatment with a preliminary cluster analysis. Therefore, these different approaches to the treatment of tinnitus can represent a fundamental difference in the neural mechanisms [7]. However, in some situations, a single hypothesis cannot accurately explain the cause of tinnitus. Moreover, although tinnitus affects large numbers of people and reduces their quality of life, evidence-based, multidisciplinary clinical practice guidelines are not yet clear [1]. Despite hearing loss being the most common cause of objective tinnitus, not enough studies to date have assessed whether variations in the characteristics of tinnitus are dependent on the type of hearing loss [5] [6].

Primary and secondary outcome measures: The key predictor variables included demographic, tinnitus, hearing-related and treatment-related variables as well as clinical factors (e.g., anxiety, depression, insomnia), which can have an impact on the treatment outcome. A 13-point reduction in Tinnitus Functional Index (TFI) scores has been defined as a successful outcome [8].

3 Method

The data were analyzed using multiple linear regression and K-MEAN regression techniques. Multiple regressions were used to establish correlations between several predictors and the response. The regression model describes the relationship between a dependent variable and many independent variables. The dependent variable in this study is the Tinnitus Reduction (after subtracting the post TFI score from the pre-TFI score), and the independent variables are group, age, gender, Duration of tinnitus, HHI Score, PHQ, and various other factors. Along with the P-value, the coefficient of determination R^2 was used to show how much of the total variation was explained by the independent variables and to identify the hypothesis that was most likely to be validated. Multiple matrices including adjusted R^2, Akaike Information Criterion (AIC), and Schwarz Bayesian Information (BIC) were used to select the best model that could be fitted based on the principle of parsimony. Following the estimation of the fitted model parameters, a diagnostic analysis was performed to determine whether the residual of the model had a normal distribution, constant variance, and zero mean. The association between these factors was investigated using the Python.

4 Analysis and Discussion

4.1 Exploratory Data Analysis

Variable	count	mean	std	min	25%	50%	75%	max
HHIScore	142	17.788732	11.370505	0	8	18	26	40
GAD	142	7.478873	5.583306	0	3	6	11	21
PHQ	142	8.028169	5.670558	0	4	7	11	27
ISI	142	12.957746	7.041793	0	8	13	18	27
SWLS	142	20.316901	7.356543	5	14	20	26	35
Hyperacusis	142	19.042254	8.496034	1	13	18.5	25	42
CFQ	142	40.591549	15.962347	7	29.25	41	50	86
Gender	142	1.436620	0.497722	1	1	1	2	2
Age	142	55.450704	12.883316	22	46.25	58	65	83
Durationoftinnitus.years	142	11.990845	12.503627	0.3	3	10	15	55
PreTFIScore	142	59.374648	18.251169	24.4	46.8	58.6	73.6	97.2
PostTFIScore	125	50.516800	21.876957	4	32	57	66	88.4

Table 1: Summary Statistics

The Table 1, above summarizes and provides information about the sample data. It informs us about the values in the data set. This includes the mean, standard deviation, minimum and maximum number of individual variable, as well as the 25th and 75th percentiles.

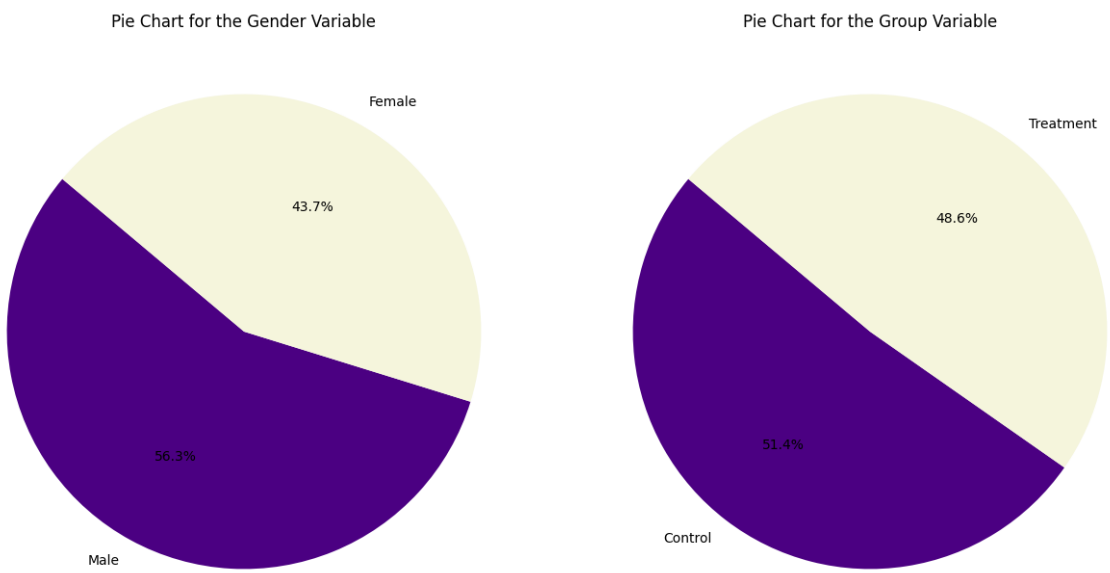


Figure 1

Figure 1 shows a pie chart of the categorical variables Gender and Group. The Gender pie chart shows that 56.3% of male subjects had Tinnitus, while 43.7% of female participants had Tinnitus. This indicates that the number of male subjects with Tinnitus is higher than the number of female subjects. However, the Group pie chart has shown that 51.4% of the subjects were assigned to the control group, while the remaining 48.6% were assigned to the treatment group.

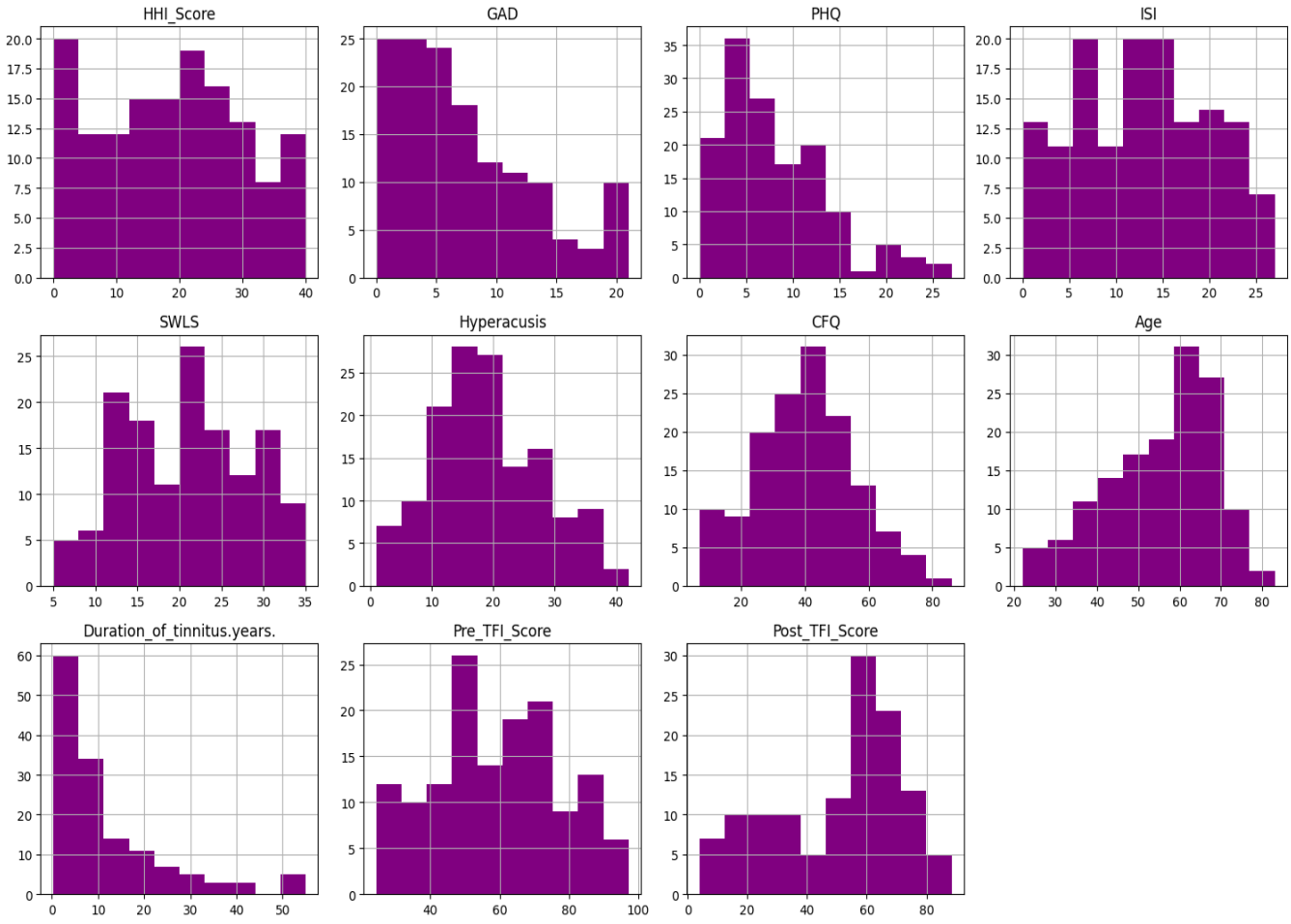


Figure 2: Subplots of the features of the Data set

Figure 2 indicates that the variables "HHI Score," "Hyperacausis," "CFQ," and "SWLS" appear to be normally distributed. However, variables "GAD," "PHQ," and "Duration of tinnitus.years." show a right-skewed pattern, whereas "Age" shows the opposite (This indicates that the majority of participants were elderly, as the age range in the histogram ranges from 22 to 83). However, there were some young ones who, according to the data, fall within the range. This indicated that Tinnitus was most likely to affect people aged 60 and up.

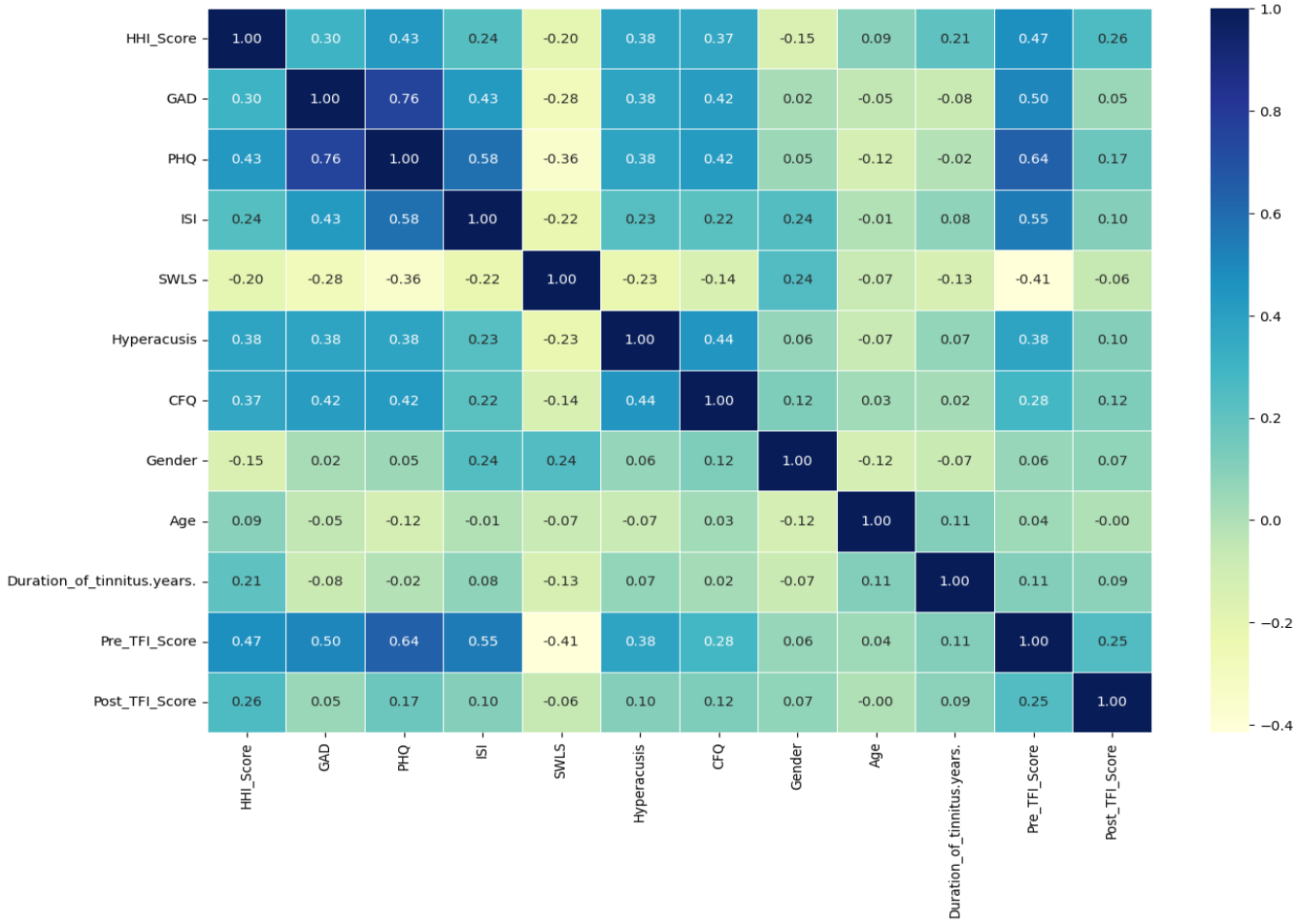


Figure 3: Correlation Matrix

Figure 3 depicts a correlation plot that proves some degree of multicollinearity in the data. For instance, "PHQ" was correlated with "GAD," "PHQ," and "Pre TFI Score." The rest of the data appears to be in essence. PHQ and GAD have a 0.76 correlation, indicating a positive relationship. This suggests that as PHQ increases, so will GAD. Moreover, PHQ and Pre TFI Score have a 0.64 correlation. This means that as PHQ increases, so will the Pre TFI Score.

4.2 Checking for Missing Values

The data was examined using the isnan syntax in Python and discovered that there was only one variable in the data with missing values. The "Post TFI Score" has 17 missing cases . There were no other anomalies in the remaining variables that we discovered. As a result, I used MICE data imputation to replace the missing values with mean values. After that, I calculated a new variable called "TFI Reduction" by subtracting "Post TFI Score" from "Pre TFI Score" .We now have a complete, ready-to-use data set.

4.3 Data Partitioning

The data contains 142 observations, and we used 80% of them to create a training dataset called "train_partition" (contains 114 observations). With the remaining 28 observations, we created a test data set called "test_partition."

4.3.1 Modeling

Fitting Forward Regression Model The assumption here is to fit a linear model to the values of the independent variables, then use Forward regression to select a subset of the variables that best fit the data (based on higher Adj.R2, least AIC, and BIC).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          TFI_Reduction   R-squared:                       0.463
Model:                            OLS   Adj. R-squared:                  0.404
Method:                 Least Squares   F-statistic:                     7.909
Date:                Thu, 29 Feb 2024   Prob (F-statistic):           9.15e-10
Time:                        16:07:56   Log-Likelihood:                -481.21
No. Observations:                 113   AIC:                             986.4
Df Residuals:                     101   BIC:                             1019.
Df Model:                          11
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          -10.6632     12.063     -0.884      0.379     -34.593      13.267
HHI_Score                       -0.0839      0.192     -0.437      0.663      -0.465       0.297
GAD                              0.3285      0.493      0.666      0.507      -0.650       1.307
PHQ                              0.5033      0.576      0.873      0.385      -0.640       1.647
ISI                              1.0226      0.319      3.205      0.002       0.390       1.655
SWLS                            -0.3211      0.273     -1.176      0.242      -0.863       0.221
Hyperacusis                      0.2380      0.243      0.981      0.329      -0.243       0.719
CFQ                             -0.1137      0.135     -0.840      0.403      -0.382       0.155
Age                             -0.0492      0.143     -0.344      0.731      -0.333       0.235
Duration_of_tinnitus.years.      0.0270      0.144      0.187      0.852      -0.260       0.314
GroupTreatment_Treatment        25.7306      3.584      7.179      0.000      18.621      32.841
Gender_2                        -3.2630      3.925     -0.831      0.408     -11.049       4.523
==============================================================================
Omnibus:                        1.873   Durbin-Watson:                   1.948
Prob(Omnibus):                  0.392   Jarque-Bera (JB):                1.923
Skew:                          -0.289   Prob(JB):                        0.382
Kurtosis:                       2.725   Cond. No.                         569.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                                 0         1
const                      -34.593307  13.266822
HHI_Score                   -0.464894   0.297061
GAD                         -0.650125   1.307223
PHQ                         -0.640215   1.646887
ISI                          0.389737   1.655377
SWLS                        -0.862837   0.220597
Hyperacusis                 -0.243230   0.719203
CFQ                         -0.382089   0.154668
Age                         -0.332976   0.234505
Duration_of_tinnitus.years. -0.259617   0.313527
GroupTreatment_Treatment    18.620545  32.840594
Gender_2                   -11.048806   4.522753
```

According to the above table, the two variables that require the most attention for further investigation are ISI and Group Treatment. The main effects of the two variables in this finding are statistically significant at the 0.05 level, with P- values of 0.002 and 0.000, respectively. Despite the fact that the p-value for the model is significant. Furthermore, the R2 and Adj.R2- 0.4633 and 0.404 indicate the existence of multicollinearity.

OLS Regression Results

Dep. Variable:	TFI_Reduction	R-squared:	0.445
Model:	OLS	Adj. R-squared:	0.424
Method:	Least Squares	F-statistic:	21.63
Date:	Thu, 29 Feb 2024	Prob (F-statistic):	3.95e-13
Time:	17:39:31	Log-Likelihood:	-483.06
No. Observations:	113	AIC:	976.1
Df Residuals:	108	BIC:	989.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-11.1392	7.303	-1.525	0.130	-25.616	3.337
GAD	0.5834	0.342	1.708	0.090	-0.094	1.260
ISI	1.0526	0.271	3.880	0.000	0.515	1.590
SWLS	-0.4648	0.243	-1.916	0.058	-0.946	0.016
GroupTreatment_Treatment	25.2825	3.421	7.390	0.000	18.501	32.064

Omnibus:	1.953	Durbin-Watson:	1.915
Prob(Omnibus):	0.377	Jarque-Bera (JB):	1.995
Skew:	-0.282	Prob(JB):	0.369
Kurtosis:	2.676	Cond. No.	114.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The multiple linear regression equation obtained is:

$$\text{TFI\_Reduction} = -11.139 + 0.583(\text{GAD}) + 1.052(\text{ISI}) - 0.465(\text{SWLS}) + 25.28(\text{GroupTreatment})$$

The final summary of the best subset possible based on adjusted R^2, residual sum of squares (R^2), Bayesian Information Criterion (BIC), and Akaike information criterion (AIC). As the table above shows, our best subset model will include GAD, ISI, SWLS, and Group Treatment. The forward process, however, kept "GAD," "ISI," and "SWLS" as significant variables because their intervals include zero and the ANOVA for parameter estimation indicates they are significant as well. The MSE 335.3753 was obtained after training the model with these variables and testing it on train data.

4.3.2 Model Diagnosis and Testing for Normality.

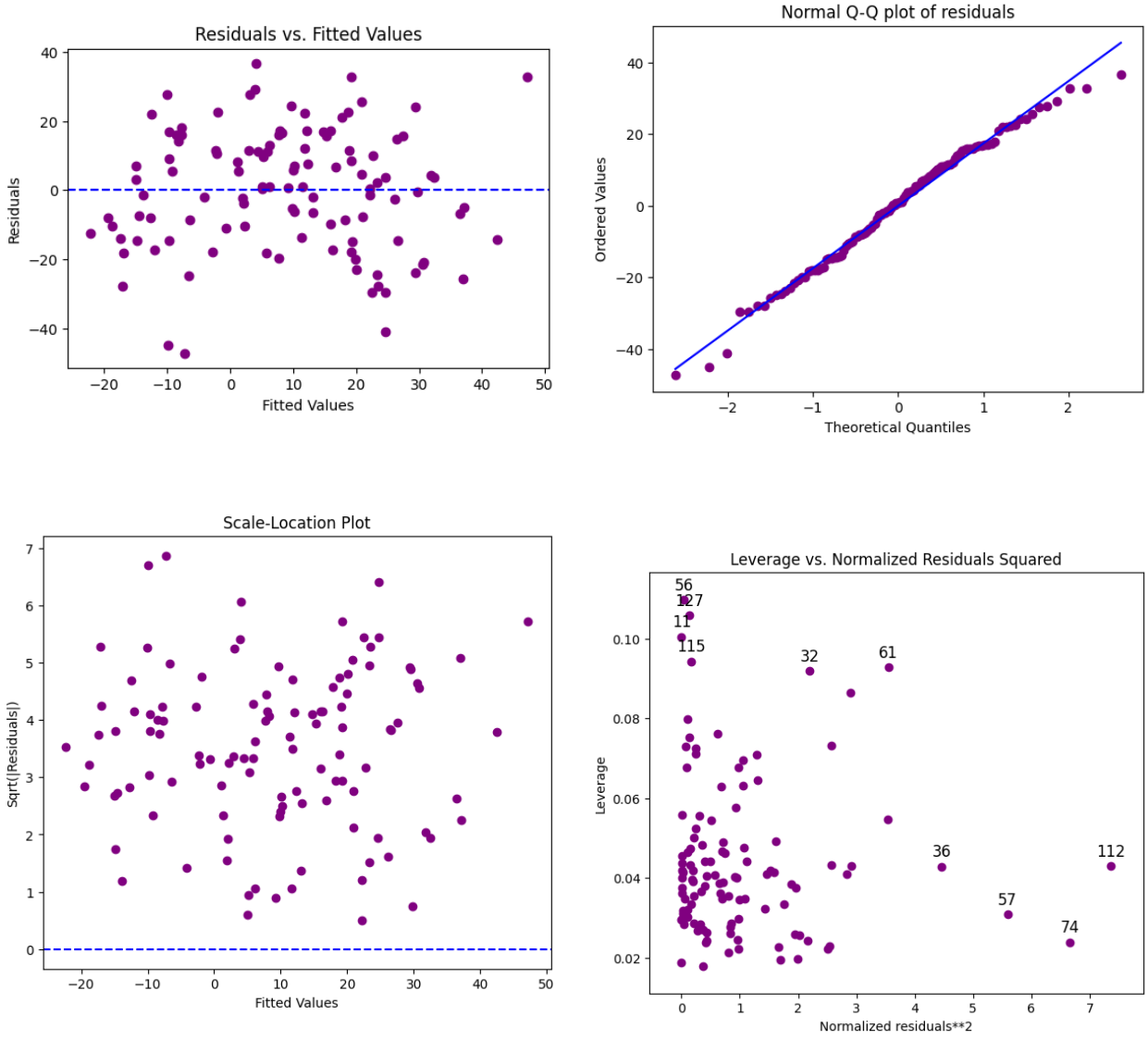


Figure 4:

Nothing unusual appears to stand out as evidence for heteroskedasticity.
According to the figure 4, In the first Residuals vs Fitted plot, we can see that the blue dotted line does not deviate significantly from the 0 (which indicates a residual value of 0). The third plot, Scale-Location, illustrates that there are no significant cases of heteroskedasticity. We want the blue dotted line to be relatively horizontal. There is no discernible pattern in the distribution of the residual points. The Q-Q plot also shows that the majority of the data points are close to the fitted line, indicating that the residuals are assumed to be Gaussian (normal), iid (independent identically distributed), and constant variance. The normality test was performed, and no significant deviations from the independence assumption were found for the residuals.

4.4 Influencing Factors of Reduction in TFI Score

TFI_Reduction = −11.139+0.583(GAD)+1.052(ISI)−0.465(SWLS)+25.28(GroupTreatment)

We imposed multiple criteria such as Adjusted R2, Residual Sum of Squares, BIC, and AIC to achieve the factors that have the greatest influence on the reduction in TFI Score and achieved the ultimate model. The model's coefficient "ISI," "GroupTreatment" and "SWLS," were significant with the exception of "GAD," and the intercept. As a result, a unit increase "GAD" ,Group Treatment and "ISI" will cause an increased in TFI Reduction by 0.58 ,25.28 and 1.083 respectively whereas a unit increase in "SWLS" will lead to a decrease in TFI Reduction by 0.465.

4.5 Making a Prediction on the Test data set.

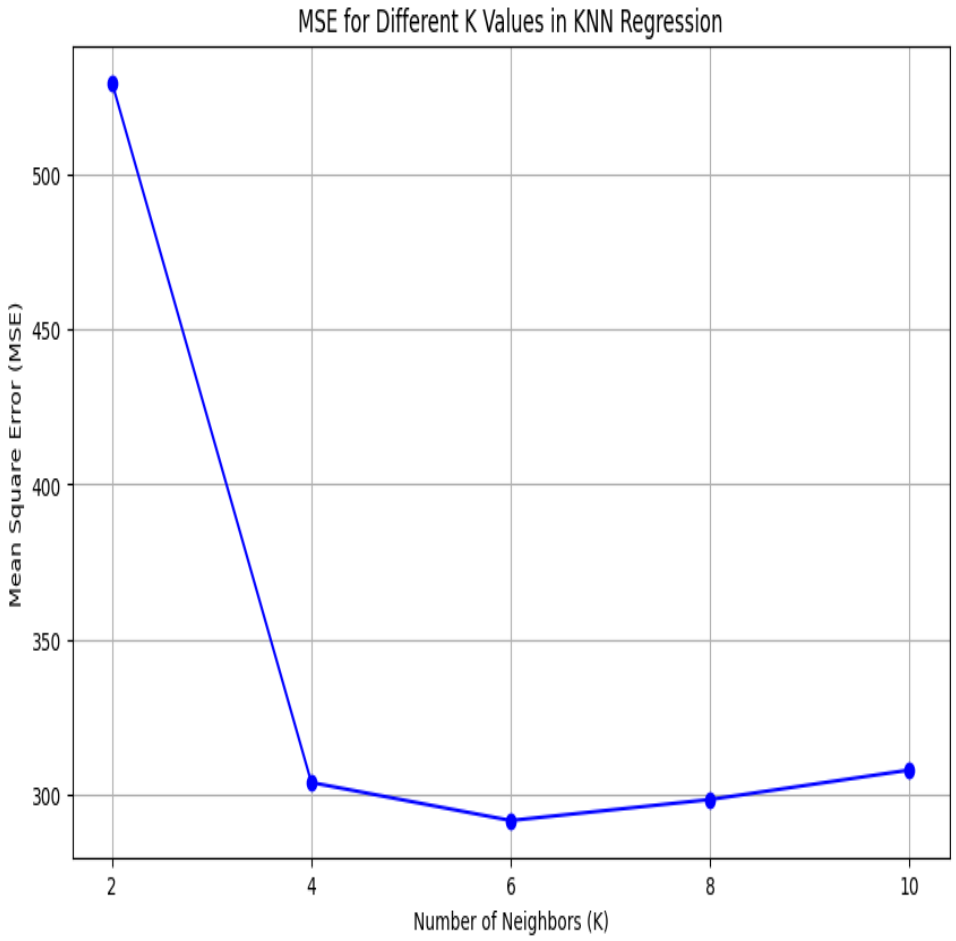
132	-2.215889
100	10.173251
94	-11.895344
41	31.691659
3	14.175958
24	15.164644
48	29.499687
96	-15.189314
2	44.556184
5	30.599068
71	-3.592538
118	3.178801
85	-8.279909
23	49.332188
92	3.179699
55	29.754580
45	33.008840
12	13.984631
59	26.588954
86	-5.378898
128	-6.526230
97	-3.497772
117	22.203907
120	13.587638
113	-4.909719
122	-3.811236
25	25.649700
44	36.882683
135	-0.417190

On the basis of the testing data, predictions were made and the estimated values were compared to the actual values. The majority of the estimates were reasonably close to the actual data. In addition, our statistical model discovered a mean Squared error amount of 335.375. Furthermore, it computes the average squared difference between observed and predicted values. When there is no error in the model, the (MSE) equals zero. As the model error increases, so does its value.

4.6 KNN Regression

Using the train data that was obtained from the partition, $K = 2, 4, 6, 8, 10$ were trained using the data partition, where 80% of the data is assigned to training and the remaining 20% to testing. Numerous different versions of K were made. The smallest of the plotted cross-validation was found to be $K = 6$; the KNN regression and plot of the K-values are shown below.

MSE for K = 2	**MSE for K = 4**	**MSE for K = 6**	**MSE for K = 8**	**MSE for K = 10**
529.643	303.869	291.613	298.384	307.898



The multiple linear regression yielded a mean squared error of 355.375, while the KNN regression resulted a mean squared error of 291.613. When we compare the two, we can see that KNN has a lower mean squared error than linear regression. And we know that when the model is error-free, the (MSE) equals zero. As the model error gets bigger, so does the value. In this study, the KNN performed better than the linear regression because the mean squared error was lower.

Appendix

Python Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib.pyplot import subplots
from functools import partial
import scipy.stats as stats
import itertools
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
from statsmodels.graphics.regressionplots import plot_leverage_resid2
import statsmodels.api as sm
from statsmodels.api import OLS
from sklearn.base import clone
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.cluster import KMeans
from sklearn.metrics import mean_squared_error
import sklearn.model_selection as skm
import sklearn.linear_model as skl
from sklearn.preprocessing import StandardScaler
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neighbors import KNeighborsRegressor

score=pd.read_csv("CaseStudy1New.csv")
score.describe()
score.drop('Unnamed: 0', axis=1, inplace=True)
score.drop('Subject_ID', axis=1, inplace=True)
score.head()
score.columns
figure, b = plt.subplots(4, 5)
figure.set_figwidth(15)
figure.set_figheight(10)
col = list(score.columns)
counter = 0
graph_names = [ 'Group',
    'HHI_Score', 'GAD', 'PHQ', 'ISI',
        'SWLS', 'Hyperacusis', 'CFQ','Gender','Age',
```

```
        'Duration_of_tinnitus.years.', 'Pre_TFI_Score', 'Post_TFI_Score'
]
for i in range(b.shape[0]):
    for j in range(b.shape[1]):
        if counter < len(col):
            b[i, j].hist(score[col[counter]], color="purple")
            b[i, j].set_title(graph_names[counter]
            if counter < len(graph_names) else col[counter])
            b[i, j].grid(True)
            counter += 1
        else:
            b[i, j].axis('off')
plt.tight_layout()
corr_matrix=score.corr()
fig,ax=plt.subplots(figsize=(15,10))
ax=sns.heatmap(corr_matrix
                ,annot=True,
               linewidths=0.5,
               fmt=".2f",
               cmap="YlGnBu")
def pie_chart(a,labels):
  groups = score[a].value_counts()
  label = groups.index.tolist()
  values = groups.values.tolist()
  label=labels
  fig =plt.figure(figsize=(8, 8))
  ax = fig.add_subplot(111)
  ax.pie(values, labels=label, autopct='%1.1f%%', startangle=140,
  colors=['indigo', 'beige'])
  ax.set_title(f'Pie Chart for the {a} Variable')
pie_chart("Gender",["Male","Female"])
pie_chart("Group",["Control","Treatment"])
score.isna().sum()
df_copy = score.copy()
group_dummies = pd.get_dummies(df_copy['Group'], prefix='GroupTreatment',
drop_first=True)
gender_dummies = pd.get_dummies(df_copy['Gender'], prefix='Gender',
drop_first=True)
df_copy = pd.concat([df_copy, group_dummies, gender_dummies], axis=1)
df_copy.drop(columns=[ 'Group', 'Gender'], inplace=True)
imputer = IterativeImputer(max_iter=5, random_state=0)
df_imputed = pd.DataFrame(imputer.fit_transform(df_copy), columns=
df_copy.columns)
missing_values_after_imputation = df_imputed.isna().sum()
df_imputed['TFI_Reduction'] = df_imputed['Pre_TFI_Score'] - df_imputed
```

```
['Post_TFI_Score']
X = df_imputed.drop(columns=['TFI_Reduction', 'Pre_TFI_Score',
'Post_TFI_Score'])
y = df_imputed['TFI_Reduction']
df_imputed
df_imputed.isna().sum()
df_imputed
seed = 2
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=seed)
X_train = sm.add_constant(X_train)
model_sm = sm.OLS(y_train, X_train).fit()
summary = model_sm.summary()
confidence_intervals = model_sm.conf_int()
print(summary)
print(confidence_intervals)
lr = LinearRegression()
sfs = SFS(lr,
          k_features='best',
          forward=True,
          floating=True,
          scoring='r2',
          cv=10)
sfs = sfs.fit(X, y)
selected_features = X.columns[list(sfs.k_feature_idx_)]
print('Selected features:', selected_features)
X_selected = X[selected_features]
X_train_selected = X_train[selected_features]
X_train_selected_sm = sm.add_constant(X_train_selected)
model_selected_sm = sm.OLS(y_train, X_train_selected_sm).fit()
summary_selected = model_selected_sm.summary()
print(summary_selected)
fitted_vals = model_selected_sm.predict()
resids = model_selected_sm.resid
"""1. Residuals vs. Fitted Values
(to check for homoscedasticity and linearity)
"""
def scatter_plot(x_label,y_label,title,data):
  plt.figure(figsize=(6, 4))
  plt.scatter(fitted_vals, data,color='purple')
  plt.axhline(y=0, color='blue', linestyle='--')
  plt.xlabel(x_label)
  plt.ylabel(y_label)
  plt.title(title)
scatter_plot("Fitted Values","Residuals",
```

```
"Residuals vs Fitted Values",resids)
"""2. Q-Q Plot (to check the normality of residuals)
"""
resids_sorted = np.sort(resids)
norm_quantiles = stats.norm.ppf((np.arange(len(resids)) + 0.5) / len(resids))
plt.scatter(norm_quantiles, resids_sorted, color='purple')
slope, intercept = np.polyfit(norm_quantiles, resids_sorted, 1)
plt.plot(norm_quantil
es, intercept + slope * norm_quantiles, 'b', color='blue')
plt.title('Normal Q-Q plot of residuals')
plt.xlabel('Theoretical Quantiles')
plt.ylabel('Ordered Values')
"""3. Scale-Location Plot (to check homoscedasticity)
"""
scatter_plot("Fitted Values","Sqrt(|Residuals|)","Scale-Location Plot"
,np.sqrt(np.abs(resids))
"""4. Leverage Plot (to identify influential cases)
"""
fig, ax = plt.subplots(figsize=(12, 10))
fig = plot_leverage_resid2(model_selected_sm, ax=ax,color='purple')
plt.title('Leverage vs. Normalized Residuals Squared')
"""Making a Prediction on the Test data set."""
X_test_selected = X_test[selected_features]
X_test_selected_sm = sm.add_constant(X_test_selected)
y_pred = model_selected_sm.predict(X_test_selected_sm)
y_pred
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
print(f"R-squared: {r2}")
"""K-Mean Regression
Question 10: Using K-NN Regression with Multiple K Values
"""
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_train_scaled, X_test_scaled,
y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,
random_state=42)
k_values = [2, 4, 6, 8, 10]
knn_models = {}
scores = {}
for k in k_values:
    knn = KNeighborsRegressor(n_neighbors=k)
    knn.fit(X_train_scaled, y_train)
    knn_models[k] = knn
```

```
    y_pred = knn.predict(X_test_scaled)
    scores[k] = mean_squared_error(y_test, y_pred)
print("KNN regression models have been fitted and evaluated for each
specified K value.")
"""Question 11: Making Predictions on the Testing Dataset and
Obtaining the Mean Square Error
"""
mse_for_each_k = {}
for k in k_values:
    # Use KNeighborsRegressor for regression
    knn_regressor = KNeighborsRegressor(n_neighbors=k)
    knn_regressor.fit(X_train_scaled, y_train)
    # Predicting the target values for the test set
    predictions = knn_regressor.predict(X_test_scaled)
    # Calculating the mean squared error (MSE) for the predictions
    mse = mean_squared_error(y_test, predictions)
    mse_for_each_k[k] = mse
    print(f"MSE for K={k}: {mse}")
# Plotting the MSE for different values of K to find the best K
plt.figure(figsize=(10, 6))
plt.plot(list(mse_for_each_k.keys()), list(mse_for_each_k.values()), '-o',
color='blue')
plt.title('MSE for Different K Values in KNN Regression')
plt.xlabel('Number of Neighbors (K)')
plt.ylabel('Mean Square Error (MSE)')
plt.xticks(list(mse_for_each_k.keys()))
plt.grid(True)
plt.show()
"""Question 12: Selecting the Optimal Number of Clusters Based on the Lowest
"""
best_k = min(mse_for_each_k, key=mse_for_each_k.get)
best_mse = mse_for_each_k[best_k]
print(f"The best K is {best_k} with the lowest test MSE of: {best_mse}")
"""Question 13:Comparison of Regression Models: Evaluating the Efficacy of
Multiple Linear Regression vs. K-NN Regression Based on Test MSE"""
mse_linear_regression = 335.37534277618977
best_k = min(mse_for_each_k, key=mse_for_each_k.get)
best_mse_kNN = mse_for_each_k[best_k]
print(f"MSE from Multiple Linear Regression: {mse_linear_regression}")
print(f"Best MSE from K-NN Regression (K={best_k}): {best_mse_kNN}")
if mse_linear_regression < best_mse_kNN:
    print("Multiple Linear Regression yields a lower MSE and is the better
    model for the dataset.")
elif mse_linear_regression > best_mse_kNN:
```

```
    print("K-NN Regression yields a lower MSE and is the better
    model for the dataset.")
else:
    print("Both models result in the same MSE, indicating equal
    performance on the dataset.")
```

References

1. Tunkel, D. E., Bauer, C. A., Sun, G. H., Rosenfeld, R. M., Chandrasekhar, S. S., Cunningham, E. R., Archer, S. M., Blakley, B. W., Carter, J. M., Granieri, E. C., Henry, J. A., Hollingsworth, D., Khan, F. A., Mitchell, S., Monfared, A., Newman, C. W., Omole, F. S., Phillips, C. D., Robinson, S. K., Taw, M. B., Tyler, R. S., Waguespack, R. & Whamond, E. J. Clinical practice guideline. OtolaryngologyHead and Neck Surgery 151, S1–S40 (2014).

2. Coelho, C. B., Santos, R., Campara, K. F. & Tyler, R. Classification of tinnitus. Otolaryngologic Clinics of North America 53, 515–529 (2020).

3. LaMarte, F. P. & Tyler, R. S. Noise-induced tinnitus. AAOHN Journal 35, 403–406 (1987).

4. Paul, W. F., Bruce, H., Valerie, J., John, K., Richardson, M. & others. Cummings otolaryngology: Head and neck surgery. Los Angeles: Mosby 2674 (2010).

5. Preece, J. P., Tyler, R. S. & Noble, W. The management of tinnitus. Geriatrics and Aging 6, 22 (2003).

6. Møller, A. R. The role of neural plasticity in tinnitus. Progress in brain research 166, 37–544 (2007).

7. Tyler, R., Coelho, C., Tao, P., Ji, H., Noble, W., Gehringer, A. & Gogel, S. Identifying tinnitus subgroups with cluster analysis. American Journal of Audiology 17, (2008).

8. Rodrigo, H., Beukes, E. W., Andersson, G., Manchaiah, V. & others. Exploratory data mining techniques (decision tree models) for examining the impact of internet-based cognitive behavioral therapy for tinnitus: Machine learning approach. Journal of medical Internet research 23, e28999 (2021).