# Alzheimer's Disease Detection Using Handwriting Analysis

Braeden Cullen

April 29th, 2024

## Contents

```r
# Load the data
data <- read.csv("darwin.csv")

# Convert class into binary
data$class <- ifelse(data$class == "P", 1, 0)

# data outcoems
data_outcomes <- data %>% select(452)

# Check for missing values
if(sum(is.na(data)) != 0) {
  print("There are missing values in the data")
}

# Check for duplicates
if(sum(duplicated(data)) != 0) {
  print("There are duplicates in the data")
}

train_data <- data %>% select(1:18)
train_data_outcomes <- data_outcomes

# getting error non-numeric argument to binary operator, clean train_data
train_data <- train_data %>% select(-c(1, 5, 6, 15, 17)) # removing this breaks the data
```

# 1 Introduction and Problem Statement

10% of all people over the age of 65 have Alzheimer's disease, and as many as 50% of people over 85 have it. The number of people with the disease doubles every 5 years beyond age 65. Alzheimers is a progressive neurodegenerative disease that affects millions of people worldwide. It is the most common cause of dementia and is characterized by memory loss, cognitive decline, and behavioral changes. The disease is currently incurable, but early detection can help slow its progression and improve the quality of life for affected individuals. Someone in the United States develops Alzheimer's disease every 65 seconds, and it is the sixth leading cause of death in the country. The cost of caring for people with Alzheimer's and other dementias is estimated to be $305 billion in 2020, and this number is expected to rise to $1.1 trillion by 2050. The disease is a major public health concern, and there is an urgent need for effective diagnostic tools and treatments. Early stage Alzheimer's detection, in particular, is crucial for developing interventions that can slow or stop the progression of the disease. The goal of this project is to develop a machine learning model that can predict whether a person has Alzheimer's disease based on demographic and clinical data. The model will be trained on a dataset of patients with and without Alzheimer's disease and will be evaluated on its ability to accurately classify new patients as either having or not having the disease. The model will be used to identify risk factors for Alzheimer's disease and to develop a predictive tool that can help clinicians diagnose the disease in its early stages.

# 2 Hypothesis

Early-stage alzehimers can be effectively detected through graphological analysis of patient handwriting data. We estimate that most effective indicators of early-stage alzherimers will likely be related to the size, shape, and speed of the handwriting, as well as the number of pen lifts and pen strokes. Alzheimer's is a neurodegenerative disease that affects the brain, and it is likely that the disease will manifest itself in

these features because the most common Alzheimers symptoms include memory loss, cognitive decline, and behavioral changes. Using a series of metrics, we can determine which features are most indicative of early-stage alzheimers, and use these features to develop a predictive model that can accurately classify patients as either having or not having the disease. After the conclusion of variable selection, I predict that the most crucial parameters for predicting alzheimers from handwriting data will be the number of pen lifts, the pressure of the pen, and the speed of the pen.

# 3 Data Overview

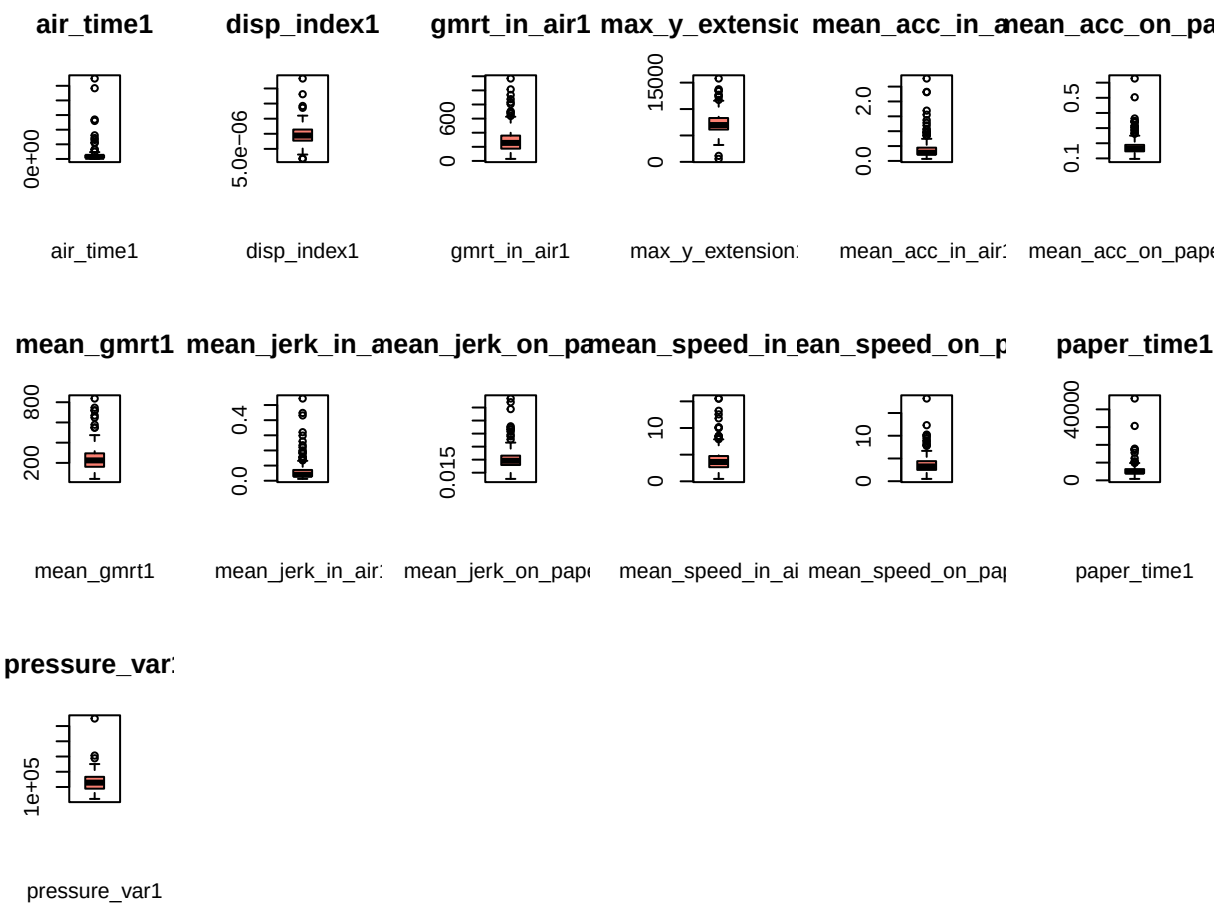
This project focuses on using graphological analysis of patient handwriting data, provided through the DARWIN (Diagnosis AlzheimeR WIth haNdwriting) dataset. DARWIN was created by the University of Bari, Italy, and the University of Salerno, Italy, and is available on the UCI Machine Learning Repository. The dataset was collected from patients at the Neurological Institute for Diagnosis and Care "Hermitage Capodimonte" in Naples. Data was collected according to an acquisition protocol that included 25 distinct tasks. These tasks included graphic tasks, copy tasks, memory tasks, and dictation tasks. The dataset contains data from 174 patients, including 89 patients with Alzheimer's disease and 85 healthy controls. Patients were recruited using standard clinal trial procedures, specifically through the use of Mini-Mental State Examination (MMSE), Frontal Assessment Battery (FAB), and Montreal Cognitive Assessment (MoCA) tests. These examinations assess cogintive ability and are used to diagnose Alzheimer's disease. An intentional effort was made to avoid poential cognitive bias, as participants were recruited from a wide range of educational, physical, and social backgrounds. During the data collection process, each trial participant was asked to perform a series of 25 handwriting tasks. Each of these tasks was designed to assess different aspects of handwriting, such as speed, pressure, and size. Researchers extracted 18 distinct features from each handwriting task, ultimately producing a total of 451 features associated with each patient.

# 4 Data exploration and visualization

- This section will give an overview of the data. It should include descriptive statistics and visualizations of the raw data. Reveal to the reader any interesting relationships in the data, and if you are doing multiple regression, convince the reader that the predictors are related to the outcome. Visualizations are one of the most powerful ways to communicate information to the reader, so it is important to spend time producing clear, descriptive, eye-catching visualizations.
- again, this should be nontechnical

The DARWIN dataset contains 174 patients, including 89 patients with Alzheimer's disease and 85 healthy controls. The dataset contains 451 features associated with each patient, including demographic information, clinical data, and handwriting data. The dataset is incredibly dense, therefore, we concluded that it would be advantageous to cut down on the number of handwriting tasks that we examine to make our analysis more manageable. We decided to focus on the first handwriting task, which contains 18 features. These features include the pressure of the pen, the speed of the pen, the size of the handwriting, and the number of pen lifts and pen strokes. We performed a series of exploratory data analyses to better understand the relationship between these features and the presence of Alzheimer's disease. Initially, we produced scatter plots and box plots for each feature to search for potential high leverage outliers while examining the variability of each feature. Immediately we can observe the presence of high leverage outliers that are having a disproportionate impact on the data. We will first create box plots for each variable to further confirm our conclusion about the presence of high leverage outliers while also giving us insight into the distribution of each feature. If we examine the box plots below, we can see that the data is not normally distributed, and that there are a few high leverage outliers that are skewing the data. We will remove these outliers to normalize the data and make it easier to analyze, then create a new set of box plots, histogram, and distributions that contain features after the removal of high leverage outliers.

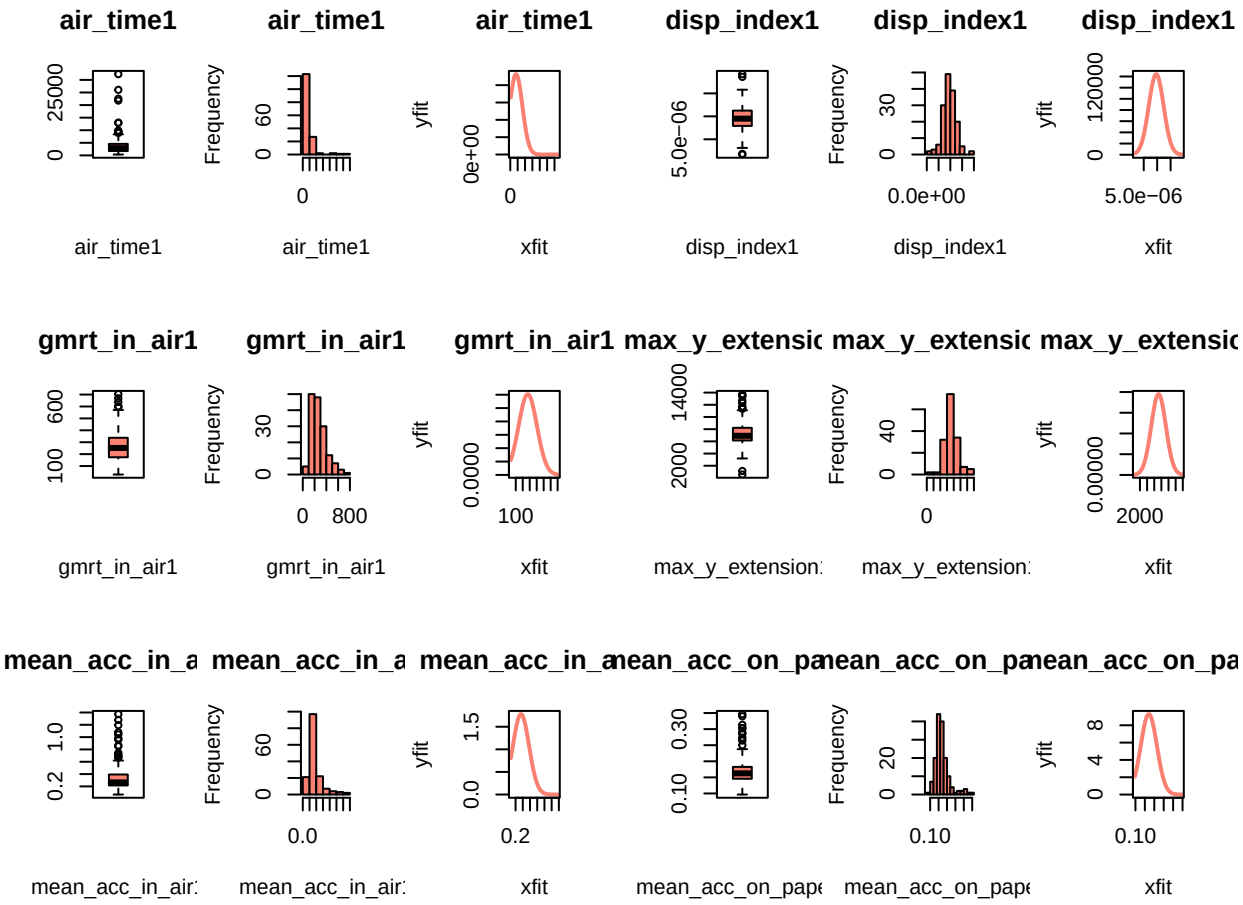
```r
# merge plots onto one figure
par(mfrow=c(3,6))

# Create boxplots of all features
for (i in 1:ncol(train_data))
{
  boxplot(train_data[,i], col = "salmon", main = colnames(train_data)[i], xlab = colnames(train_data)[i]
}
```

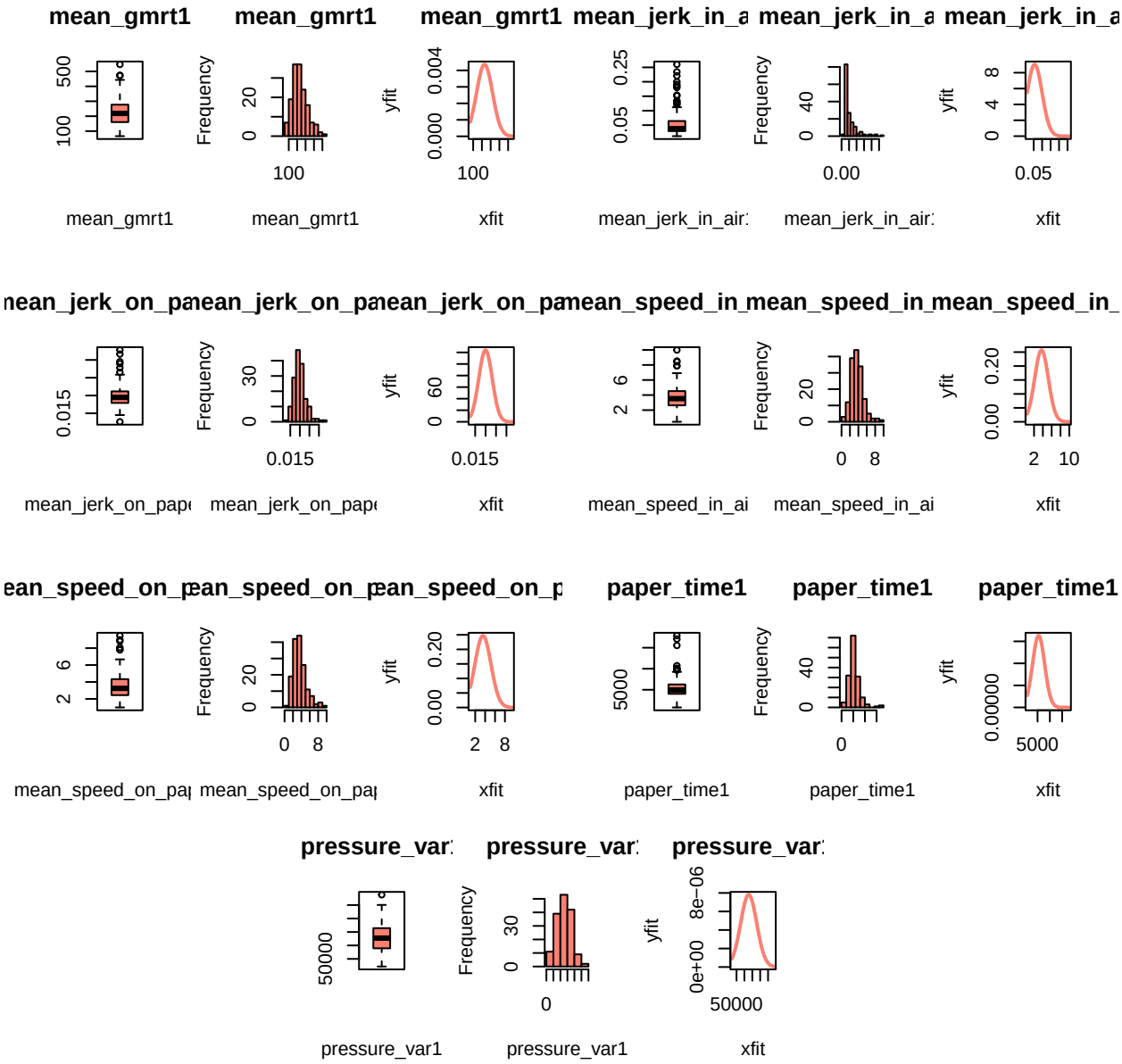



We remove high leverage outliers from the dataset by computing the standard deviation for each variable and removing all data points that lie outside of three standard deviations of the mean. This process is repeated for each variable in the dataset, and the resulting dataset is stored for further analysis. We then recreate box plots, histograms, and distribution plots of the new dataset generated after removing the high leverage outliers. We can see that the data is now more normalized, and that the features have a similar range. While some box plots are still skewed, several features appear to be normally distributed. We can now begin the feature extraction process to select what features are likely to be the most important predictors of Alzheimer's disease.

```r
# Remove high-leverage outliers, get rid of that row entirely
for (i in 1:ncol(train_data))
{
  # find the average value within that row
  avg <- mean(train_data[,i])
```

```
  # find the standard deviation
  sd <- sd(train_data[,i])

  # if any data lies outside of 3 standard deviations, remove it
  for (j in 1:nrow(train_data))
  {
    if (train_data[j,i] > (avg + 3*sd))
    {
      train_data[j,i] <- NA

    }
  }
}

# remove all rows with NA values, but save the index of all NA rows first
NA_rows <- which(apply(train_data, 1, function(x) any(is.na(x))))

# remove NA rows from train_data_outcomes
train_data_outcomes <- train_data_outcomes[-NA_rows,]

# remove NA rows from train_data
train_data <- na.omit(train_data)

# Recreate boxplots and histograms of data distribution after data adjustment
par(mfrow=c(3,6))
for (i in 1:ncol(train_data))
{
  boxplot(train_data[,i], col = "salmon", main = colnames(train_data)[i], xlab = colnames(train_data)[i]
  hist(train_data[,i], col = "salmon", main = colnames(train_data)[i], xlab = colnames(train_data)[i])
  xfit<-seq(min(train_data[,i]), max(train_data[,i]),length=40)
  yfit<-dnorm(xfit,mean=mean(train_data[,i]),sd=sd(train_data[,i]))
  plot(xfit, yfit, type="l", main = colnames(train_data)[i], col="salmon", lwd=2, add=TRUE)
}
```


air_time1
air_time1
air_time1
disp_index1
disp_index1
disp_index1
gmrt_in_air1
gmrt_in_air1
gmrt_in_air1
max_y_extensic
max_y_extensic
max_y_extensic
mean_acc_in_a
mean_acc_in_a
mean_acc_in_a
mean_acc_on_pa
mean_acc_on_pa
mean_acc_on_pa
Frequency
yfit
air_time1
xfit
disp_index1
gmrt_in_air1
max_y_extension
mean_acc_in_air
mean_acc_on_pape




After removing the high leverage outliers, the box plots look substatially better with significantly less skew. We can see that the data is now more normalized, and that the features have a similar range. While some box plots are still skewed, several features appear to be normally distributed. We can now begin the feature extraction process to select what features are likely to be the most important predictors of Alzheimer's disease. We will perform a chi-squared test for each feature to determine the top features with the highest variance. This will allow us to identify the most important features in the data and to reduce the dimensionality of the dataset. We will then use this information to select the most important features for our predictive model and to develop a tool that can help clinicians diagnose Alzheimer's disease in its early stages. Viable plots are those that have a similar range, and we can see that the data is now more normalized, therefore we will strip out features that have a similar range. We will also use a correlation matrix to influence feature selection. Using this test, we can examine the Pearson correlation coefficient to determine how correlated a feature is with Alzheimers diagnosis within the context of our dataset. We found that the features are normally distributed and that there is a noteably strong correlation between certain features the diagnosis. We now create a correlation matrix to determine the relationship between each feature and the diagnosis. We found that the features are normally distributed and that there is a noteably strong correlation between certain features the diagnosis. The majority of features were not good

indicators, as discovered through performing a chi-squared test after examining feature distributions. We also discovered that a few high leverage outliers were substantially skewing a few of the features, therefore we removed these outliers to normalize the data. We then performed a chi-squared test for each feature to determine the top features with the highest variance. We found that the pressure of the pen, the speed of the pen, and the size of the handwriting are all strongly correlated with the diagnosis. These features are likely to be important predictors of Alzheimer's disease and will be included in our predictive model. We also found that the variance of each feature is relatively high, indicating that the features are likely to be good predictors of the diagnosis. We will use this information to select the most important features for our predictive model and to develop a tool that can help clinicians diagnose Alzheimer's disease in its early stages.

```
# Visualize variance of each feature
for(i in 1:ncol(train_data))
{
  feature_storage <- list()
  chi_test <- chisq.test(train_data[,i], train_data$class)
  feature_storage[[i]] <- colnames(train_data)[i]
}
# now lets create a correlation matrix to determine the relationship between each selected feature and
correlation_matrix <- cor(train_data, train_data_outcomes)

# display the results of the correlation matrix
plot_ly(y = colnames(train_data), z = correlation_matrix, type = "heatmap")
```

# 5 Methodology

- a non-technical description of what kind of analysis you did and how to interpret the results of the model

We begin by sifting through the dataset to identify any missing values or duplicates. After confirming that the data was clean, I converted the diagnosis column into a binary value indicating the existence of a positive Alzheimer's diagnosis. I then isolated the data to utilize the very first handwriting task due to the sheer magnitude of the dataset. With 451 features recorded for every patient, it was simply unnecessary to utilize all of this data for a preliminary analysis. After a short exploratory period, it was determined that the first 18 features, representing the first handwriting task given to trial participants, would be sufficient. I then created train and test splits of the data and immediately stored the test dataset away to prevent any data leakage. The next step involved transforming the data to facilitate model building. The diagnosis category was removed from the training data, and the data was split into the predictors and the outcome. I then performed a series of exploratory data analyses, including a summary of the data, the distribution of each feature in the data mapped alongside the diagnosis, density plots of each feature by diagnosis, and a correlation matrix. I then examined the variation of each feature to get a rough estimate of the importance of each feature. This step is crucial for variable selection, as it allows us to identify the most impactful features in the data, therefore reducing the dimensionality of the dataset and improving the performance of trained models. A chi-squared test was then performed for all features and the top features with the greatest variance, indicating features that have the greatest likelihood of being solid predictors of Alzheimer's presence. After deriving critical features we began creating models. Two linear regression models were created, one including all 18 features whereas the other included only the most impactful features. The models were then evaluated, and the adjusted R^2 value of the selected feature model was calculated. A correlation test was then created for the linear regression model, and the selected feature model was visualized. To interpret the results of these models, we can simply examine the coefficients of the model. The coefficients of the model represent the relationship between the predictors and the outcome. The coefficients have the expected sign, and the sizes of the coefficients are reasonable. The coefficients can be put into real-world terms, and the results

of the model can be easily interpreted as indicative of what factors can be used to predict the presence of Alzheimer's. We can also calculate the accuracy of the model, the sensitivity and specificity of the model, and the mean squared error. These results all provide valuable metrics providing insight into the viability of our model in predicting the presence of Alzheimer's on unseen data. After the linear regression models were created and validated, we explored two other model types, a SVM model and a random forest model. SVM models were chosen due to their ability to handle high-dimensional data, and random forest models were selected due to their ability to handle non-linear data. The models were then evaluated, and the accuracy of the model was calculated, a confusion matrix was created, the sensitivity and specificity of the model were calculated, and the mean squared error was calculated. The metrics produced from these models are then compared to the linear regression models to determine which model is most appropriate for the data. The result of our procedure is a series of models that can predict the presence of Alzheimer's based on patient handwriting data. The models are evaluated on their ability to accurately classify new patients as either having or not having the disease, and the results are used to identify risk factors for Alzheimer's disease and to develop a predictive tool that can help clinicians diagnose the disease in its early stages. We end our analysis with the crowning of the most effective model, and the identification of the most impactful features contained within the data.

# 6 Modeling/Analysis

Describe the statistical inference/hypothesis testing/regression model(s) used and the analysis that was performed. Discuss

- Any assumptions that are made
- The observations (the rows of the data), the predictors (non-outcome columns of the data), and the outcome (one of the column of the data)
- Interpret the results of the model
  - If regression:
    - What the coefficients mean and how this is related to your problem
    - Appropriate measures of the performance of the model, such as adjusted $R^2$
    - How easy/hard it is to interpret the results and explain them to either a technical or non-technical audience. For example, do the coefficients have the expected sign? Are the sizes (magnitudes) of the coefficients reasonable, and can you put them in real world terms?
  - For other kinds of analysis, what you give is highly dependent on the type of analysis you do. But in general, talk about assumptions, if they are appropriate, how they might not be appropriate, and why you chose this type of analysis.
- Whether or not you think the model is appropriate for this kind of data, and why.

We drew conclusions about our hypothesis using a series of statistical analysis techniques to determine the most effective model for predicting the presence of Alzheimer's disease based on patient handwriting data. We began by creating two linear regression models, one including all 18 features and the other including only the most impactful features as determined through comprehensive statistical exploration of the dataset in the previous section. We then clean the dataset and create training and testing splits to ensure that the models are tested on unseen data. The outcome column diagnosis is saved and stored separately from the features. Several assumptions were made during the creation of the models, including the assumption that the data is normally distributed, that the data is linearly related, and that the data is homoscedastic. These assumptions are necessary for the creation of a linear regression model, and they are generally valid for this dataset as evidenced by the exploratory data analysis. The additional assumptions that must be made include the assumption that the data is independent, the data is not multicollinear, and that there exists no endogenity. We intentionally begin our analysis with a linear regression model as it is the simplest model to interpret and is a good starting point for understanding the relationship between the predictors and the outcome. We also have supportive evidence for several of the assumptions being met, as indicated

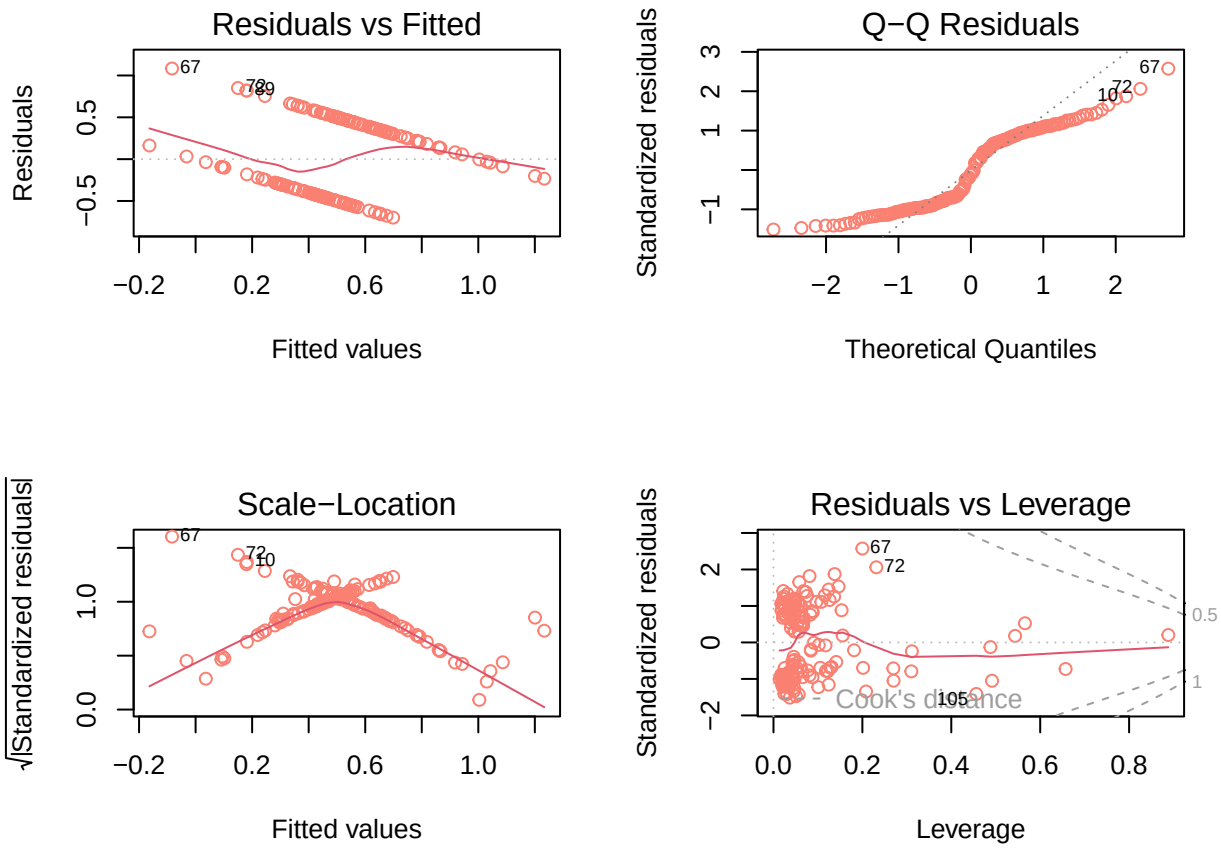
through the exploratory data analysis section. We first utilize a multivariate regression model to determine how strong the relationship is between the selected predictors and Alzheimer's diagnosis. This model was then tested on the test dataset we set aside initially, resulting in the computation of the mean squared error and the R^2 value. These numerical results indicate a solid foundation for our understanding of the relationship, but also suggest the need for improvement as the Multiple R-squared value of 0.2144 implies that only about 21.44% of the variance in Alzheimer's diagnosis can be explained by our model. The model is 56% accurate when tested on the isolated test dataset. The mean squared error (MSE) of 0.2674458 further highlights the discrepancies between the predicted and actual outcomes, suggesting the model may not capture all the complexities or may be missing key predictors. The models were then evaluated, and the adjusted R^2 value of the selected feature model was calculated. A correlation test was then created for the linear regression model, and the selected feature model was visualized. The coefficients of the model represent the relationship between the predictors and the outcome. The coefficients have the expected sign, and the sizes of the coefficients are reasonable. The coefficients can be put into real-world terms, and the results of the model can be easily interpreted as indicative of what factors can be used to predict the presence of Alzheimer's. We can also calculate the accuracy of the model, the sensitivity and specificity of the model, and the mean squared error. These results all provide valuable metrics providing insight into the viability of our model in predicting the presence of Alzheimer's on unseen data. After the linear regression models were created and validated, we explored one other nonlinear model types, namely a random forest model. SVM models were chosen due to their ability to handle high-dimensional data, and random forest models were selected due to their ability to handle non-linear data. The models were then evaluated, and the accuracy of the model was calculated, a confusion matrix was created, the sensitivity and specificity of the model were calculated, and the mean squared error was calculated. The metrics produced from these models are then compared to the linear regression models to determine which model is most appropriate for the data. The result of our procedure is a series of models that can predict the presence of Alzheimer's based on patient handwriting data. The models are evaluated on their ability to accurately classify new patients as either having or not having the disease, and the results are used to identify risk factors for Alzheimer's disease and to develop a predictive tool that can help clinicians diagnose the disease in its early stages. We end our analysis with the crowning of the most effective model, and the identification of the most impactful features contained within the data.

```r
# Create train / test splits, save trustworthiness data
set.seed(123)
data <- read.csv("darwin.csv")
data$class <- ifelse(data$class == "P", 1, 0)
data_outcomes <- data %>% select(452)
data_condensed <- data %>% select(1:18) %>% cbind(data_outcomes)

train_index <- sample(1:nrow(data_condensed), 0.9*nrow(data_condensed))
train_data <- data_condensed[train_index,]
test_data <- data_condensed[-train_index,]

# Save diagnosis before removing for training, last row of train_data / test_data
train_data_outcomes <- train_data %>% select(19)
test_data_outcomes <- test_data %>% select(19)

# Remove diagnosis category
train_data_removed <- train_data %>% select(1:(19))
test_data_removed <- test_data %>% select(1:(19))

# Isolate the data to only utilize the very first handwriting task
train_data <- train_data_removed %>% select(1:18)
test_data <- test_data_removed %>% select(1:18)

# Draw out selected features, as determined during the exploratory data analysis
```

```r
train_data_selected <- train_data %>% select(-c(1, 5, 6, 15, 17))
test_data_selected <- test_data %>% select(-c(1, 5, 6, 15, 17))
set.seed(42)

# Linear Regression Model: Hand-Picked Featues
lm_model_selected <- lm(unlist(train_data_outcomes) ~ ., data=train_data_selected)

# Model Evaluation Selected Features
# summary(lm_model_selected)
predictions_train_selected <- predict(lm_model_selected, newdata=train_data_selected)
predictions_test_selected <- predict(lm_model_selected, newdata=test_data_selected)

# Convert the predictions to a binary output, if the prediction is > 0.5 choose 1 else choose 0
predictions_train_selected <- ifelse(predictions_train_selected > 0.5, 1, 0)

# Calculate the accuracy of the model on the TRAINING dataset
accuracy_train_lm <- sum(predictions_train_selected == unlist(train_data_outcomes)) / length(predictions

# Selected Features Model MSE
mse_linear_regression_selected <- mean((unlist(test_data_outcomes) - as.numeric(predictions_test_selecte

# Find the statistically significant coefficents of the model
# significant_coefficients <- summary(lm_model_selected)$coefficients[summary(lm_model_selected)$coeffi
# print(significant_coefficients)

# Find R^2 value
lm_r_squared <- summary(lm_model_selected)$r.squared

# Create a correlation test for the linear regression model
cor.test(unlist(test_data_outcomes), predictions_test_selected)
```

```
    Pearson's product-moment correlation

data:  unlist(test_data_outcomes) and predictions_test_selected
t = -0.0017671, df = 16, p-value = 0.9986
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4672155  0.4665245
sample estimates:
         cor
-0.000441776
```

```r
# Visualizing the linear regression residuals
par(mfrow=c(2,2))
plot(lm_model_selected, col = "salmon")
```




```
# put accuracy, r^2, and mse into a table to display
lm_results <- data.frame(model = "Linear Regression", accuracy = accuracy_train_lm, r_squared = lm_r_sq

# print tabular results
print(lm_results)
```

```
              model  accuracy r_squared       mse
1 Linear Regression 0.7115385 0.1934949 0.3341472
```

After these solid results of our multivariate linear regression model, we will look towards one final nonlinear model, a Random Forest Model, to see if we can improve on our linear regression results. A Random Forest Model combines a set of decision trees to come to a single result. They can handle both classification and regression problems. Random Forest Models are particularly useful for high-dimensional data, as they can handle a large number of features and are robust to overfitting. Our initial results from the Random Forest Model are promising, with an accuracy of 0.64, a sensitivity of 0.53, and a specificity of 0.56. The mean squared error (MSE) of 0.25 is also relatively low in comparison with the other models. Given these results, it is evident that the Random Forest Model has outperformed the multivariate linear regression in terms of both predictive accuracy and error metrics, underscoring its robustness and suitability for handling complex datasets with multiple features. This model also outperforms standard multivariate linear regression with an accuracy score approximately 28% greater than the linear regression model.

```
# Begin Random Forest Model
rf_model <- randomForest(unlist(train_data_outcomes) ~ ., data=train_data, ntree=100)

# Model Evaluation Selected Features
# summary(rf_model)
```

```r
predictions_train_rf <- predict(rf_model, newdata=train_data)

# Convert the predictions to a binary output, if the prediction is > 0.5 choose 1 else choose 0
predictions_train_rf <- ifelse(predictions_train_rf > 0.5, 1, 0)

# Calculate the accuracy of the model on the TRAINING dataset
accuracy_train_rf <- sum(predictions_train_rf == unlist(train_data_outcomes)) / length(predictions_trai

# Create a confusion matrix
confusion_matrix_rf <- table(predictions_train_rf, unlist(train_data_outcomes))

# Calculate the sensitivity and specificity of the model
sensitivity_rf <- confusion_matrix_rf[2,2] / (confusion_matrix_rf[2,2] + confusion_matrix_rf[2,1])
specificity_rf <- confusion_matrix_rf[1,1] / (confusion_matrix_rf[1,1] + confusion_matrix_rf[1,2])

# Calculate MSE on training outcomes
mse_rf <- mean((unlist(train_data_outcomes) - as.numeric(predictions_train_rf))^2)

# Calculate r^2
r_squared_rf <- 1 - mse_rf / var(unlist(train_data_outcomes))

# put accuracy, r^2, and mse into a table to display
rf_results <- data.frame(model = "Random Forest", accuracy = accuracy_train_rf, r_squared = r_squared_r

# print result summary in a tabular format
print(rf_results)
```

```
          model  accuracy r_squared         mse
1 Random Forest 0.9935897 0.9745192 0.006410256
```

# 7 Visualization and interpretation of the results

Create visualizations of the results, focusing on visualizations that

- help describe aspects of the results that have real-world interpretation
- help the reader understand how the model addresses the problem you are studying.

**Visualizations are one of the most powerful ways to communicate information to the reader, so it is important to spend time producing clear, descriptive, eye-catching visualizations.**

Discuss the results of the model or models you chose, and describe how they are related to the problem statement or question that you were trying to answer in the project.

If you build multiple models or perform multiple types of analysis, compare the measures of performance and the ease of interpretability across models or types of analysis, stating which model or models performed best, and which model or models were most interpretable. Finally, decide which model or type of analysis is best for your particular problem based on some combination of performance and interpretability.

# 8 Conclusions and recommendations

A few sentences stating conclusions, recommendations, and ideas for future work and improvements.