

# ST447 Final Project : Burgess Hill v Wood Green

## Data Analysis and Statistical Methods

Candidate Number: 37984

2024-12-06

```
ID = 202426731
XYZprofile(ID) # Age: 21, Gender: Female, Home Address: Burgess Hill
```

```
## The profile of XYZ:
## - Age: 21
## - Gender: Female
## - Home address: Burgess Hill
```

## Introduction

The choice of a driving test center can impact a candidate's chances of passing due to factors like location-specific passing rates and individual demographics. This project helps XYZ, a 21-year-old female, decide between **Wood Green** (near LSE) and the nearest center to her home in **Burgess Hill**.

To make an informed recommendation, I analyzed passing rates at both centers from 2008–09 to 2023–24, employing exploratory data visualization, logistic regression modeling, and statistical tests like the t-Test and the Wald test. These analyses provide a robust, data-driven recommendation to maximize her chances of success.

## Explorating the Data

For this analysis, we utilize data from 2008-09 to 2023-24. This decision is based on the desire to gain a comprehensive and long-term understanding of trends and patterns in passing rates at the two test centers. Including data from 2008-09 ensures a sufficiently large sample size, improving the statistical robustness of the analysis. This is particularly important when comparing centers or specific subgroups like age or gender.

```
# Reading the data
data <- read_excel("burgesshill_woodgreen_data.xlsx")
```

```
data$Year <- factor(data$Year) # Converting Year to factor
data$Area <- factor(data$Area) # Converting Area to factor

# Converting all numeric values
data[, c("Age", "Male_Conducted", "Male_Passes", "Male_PassRate",
        "Female_Conducted", "Female_Passes", "Female_PassRate",
        "Total_Conducted", "Total_Passes", "Total_PassRate")] <-
  lapply(data[, c("Age", "Male_Conducted", "Male_Passes", "Male_PassRate",
```

```

      "Female_Conducted", "Female_Passes", "Female_PassRate",
      "Total_Conducted", "Total_Passes", "Total_PassRate"]],
function(x) as.numeric(as.character(x)))

```

#### Overall Statistics for the two Test Centres:

```

##      Year      Age  Male_Conducted  Male_Passes  Male_PassRate
## 2008-09: 18  Min.   :17  Min.    : 16.0  Min.    : 6.0  Min.    :32.69
## 2009-10: 18  1st Qu.:19  1st Qu.: 140.8  1st Qu.: 69.0  1st Qu.:43.81
## 2010-11: 18  Median :21  Median : 189.5  Median : 91.0  Median :48.12
## 2011-12: 18  Mean    :21  Mean    : 251.4  Mean    :125.1  Mean    :48.76
## 2012-13: 18  3rd Qu.:23  3rd Qu.: 284.2  3rd Qu.:132.0  3rd Qu.:53.85
## 2013-14: 18  Max.    :25  Max.    :1667.0  Max.    :862.0  Max.    :65.11
## (Other):180
## Female_Conducted Female_Passes  Female_PassRate Total_Conducted
## Min.    : 14.0  Min.    : 6.0  Min.    :26.54  Min.    : 37.0
## 1st Qu.: 154.8  1st Qu.: 65.0  1st Qu.:38.45  1st Qu.: 305.0
## Median : 215.0  Median : 86.0  Median :42.51  Median : 404.0
## Mean    : 252.5  Mean    :112.7  Mean    :42.94  Mean    : 503.9
## 3rd Qu.: 285.5  3rd Qu.:115.0  3rd Qu.:47.57  3rd Qu.: 562.0
## Max.    :1330.0  Max.    :688.0  Max.    :63.64  Max.    :2997.0
##
## Total_Passes  Total_PassRate      Area
## Min.    : 15.0  Min.    :31.91  Burgess Hill:144
## 1st Qu.: 137.0  1st Qu.:41.24  Wood Green  :144
## Median : 176.0  Median :45.31
## Mean    : 237.8  Mean    :45.70
## 3rd Qu.: 245.2  3rd Qu.:49.76
## Max.    :1550.0  Max.    :63.93
##

```

```

# Splitting the dataset into subsets using Area (Burgess Hill and Wood Green)
burgess_hill_data <- subset(data, Area == "Burgess Hill")
wood_green_data <- subset(data, Area == "Wood Green")

```

The average pass rates for the two centres were calculated to provide an overall measure of their performance. The results are as follows:

```

## Overall Pass Rates of the two Test Centres:
## Burgess Hill: 48.40739 %
## Wood Green: 42.98425 %

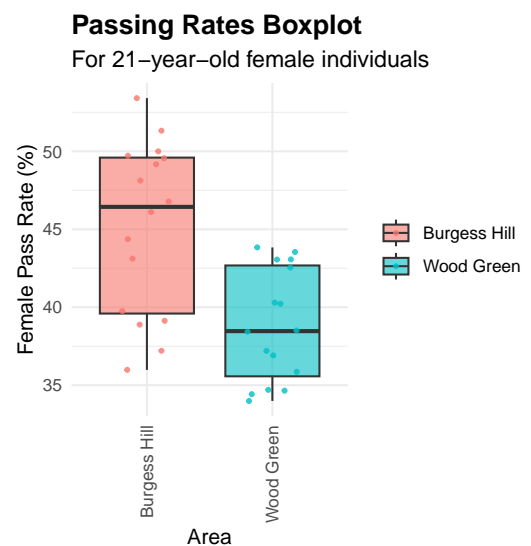
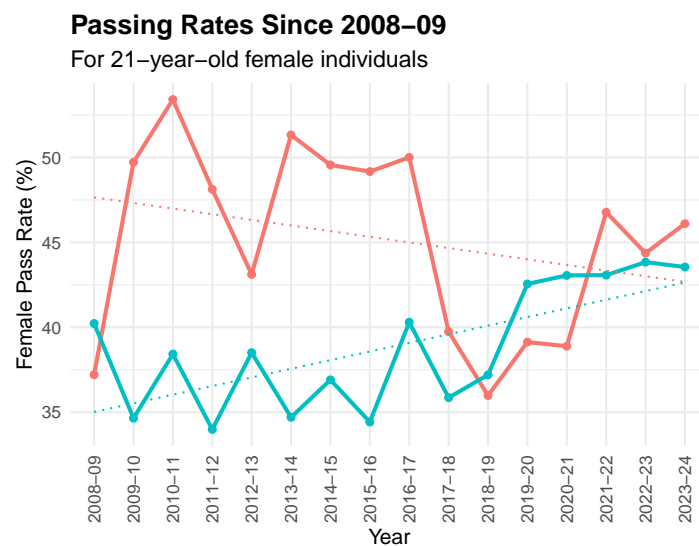
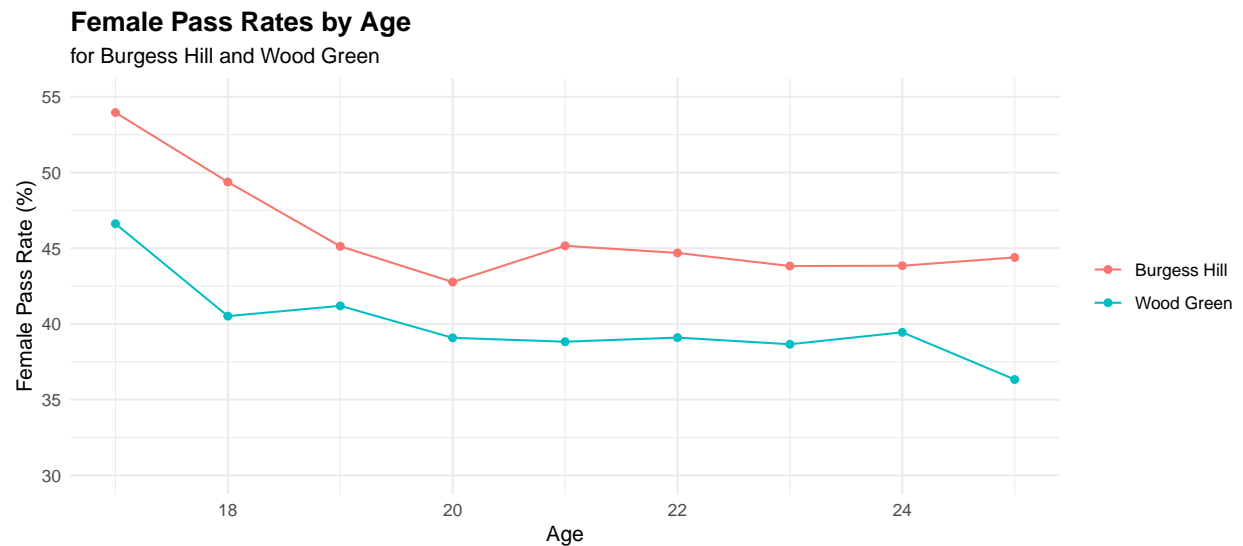
## Female Pass Rates of the two Test Centres:
## Burgess Hill: 45.90703 %
## Wood Green: 39.97704 %

## Female (Age 21) Pass Rates of the two Test Centres:
## Burgess Hill: 45.16324 %
## Wood Green: 38.82878 %

```

## Graphical Representation of the Data

Visualizations comparing female passing rates by age and over time revealed that Burgess Hill consistently outperformed Wood Green. While Burgess Hill exhibited greater variability, it maintained higher averages across all subgroups and time periods. So already we have a tendency to suggest Burgess Hill as her test centre from the figure, but we analyse the the data further.

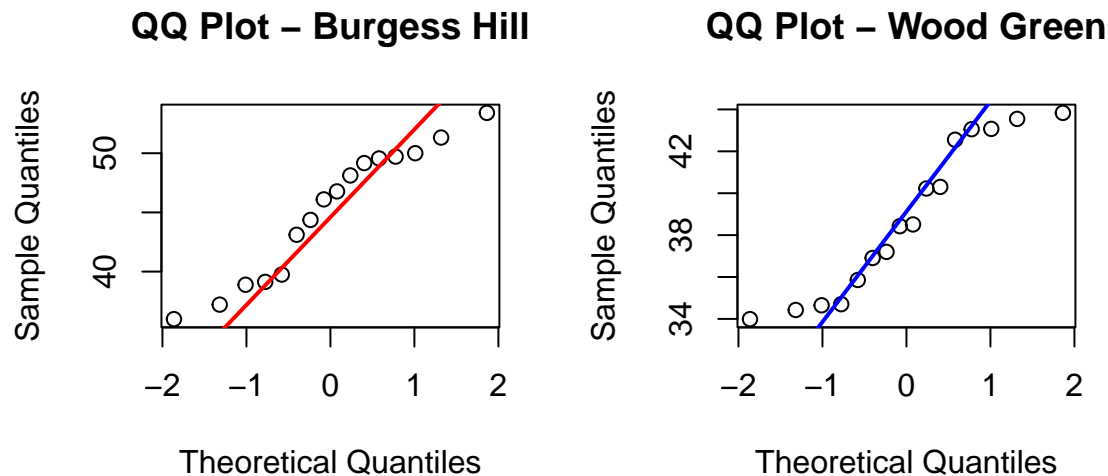


The line chart on the left compares the passing rates of 21-year-old females in Burgess Hill and Wood Green from 2008–09 to 2023–24. Unlike the previous graph, this chart highlights fluctuations over time, with Burgess Hill showing a downward trend and Wood Green displaying a slight upward trend. Despite these patterns, Burgess Hill consistently maintained higher passing rates overall.

The accompanying boxplot visually compares the average passing rates between the two centers. Burgess Hill has a higher average passing rate of 45.16% compared to Wood Green's 38.83%. However, the results from Burgess Hill exhibit greater variability.

## Checking for Normality

To perform t-Test we need to prove that the data follows a normal distribution, for small sample sizes ( $n < 30$ ), this assumption is critical.



The points generally follow the red line, indicating that the sample distribution is roughly normal, but there are deviations at the tails (extreme values). The sample distribution for Wood Green deviates more from normality compared to the left plot, with potential skewness or a different distribution shape, such as being heavy-tailed or having outliers.

It is still unclear from the QQ plots so we perform the Shapiro-Wilk test for normality. This test evaluates whether a dataset is normally distributed.

```
shapiro.test(Burgess_Hill_Age21_Female) # Normality test for Burgess Hill
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: Burgess_Hill_Age21_Female  
## W = 0.93, p-value = 0.2438
```

```
shapiro.test(Wood_Green_Age21_Female) # Normality test for Wood Green
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: Wood_Green_Age21_Female  
## W = 0.90029, p-value = 0.08124
```

For both centers, the Shapiro-Wilk test does not provide enough evidence to conclude that the data is not normally distributed. So we proceed to perform t-test to check whether there is a statistically significant difference between the means of two test centres.

## t-Test

```
# Perform a t-test
t_test_results <- t.test(Burgess_Hill_Age21_Female, Wood_Green_Age21_Female, alternative = "two.sided",
# Print the result
print(t_test_results)

##
## Welch Two Sample t-test
##
## data: Burgess_Hill_Age21_Female and Wood_Green_Age21_Female
## t = 3.8522, df = 25.819, p-value = 0.0006929
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.953263 9.715661
## sample estimates:
## mean of x mean of y
## 45.16324 38.82878
```

The Two Sample t-test indicates that the mean for Burgess Hill (45.16) is significantly higher than the mean for Wood Green (38.83). The difference is statistically significant ( $p < 0.001$ ), and we can be confident that the true mean difference is between 2.95 and 9.72. This result provides strong evidence that the higher mean observed in Burgess Hill is not due to random chance.

## Logistic Regression

Regression is a logical method for estimating the likelihood of passing the driving test. While linear regression can be used to predict passing rates with Age, Gender, and Area as predictors (treating this as a classification problem), it is not well-suited for binary outcomes. Linear regression may produce probabilities outside the valid range of 0 to 1, making it inappropriate for this context.

Logistic regression offers a more appropriate approach. By applying the logit link function, it maps a linear combination of predictors into the range (0,1). These values can then be interpreted as probabilities, representing the likelihood of passing the test given specific individual characteristics:

$$P(Y = 1|Age, Area, Gender) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 Age + \beta_2 Area + \beta_3 Gender)}}$$

Alternatively, in terms of the logit function:

$$\text{logit}(P(Y = 1|Age, Area, Gender)) = \ln \left( \frac{P(Y = 1|Age, Area, Gender)}{1 - P(Y = 1|Age, Area, Gender)} \right) = \beta_0 + \beta_1 Age + \beta_2 Area + \beta_3 Gender$$

Using logistic regression, we can model the probability of passing the driving test based on Age, Gender, and Area, offering a reliable and meaningful analysis.

Before fitting the logistic regression model, the data needs to be restructured. Specifically, it should be organized into a matrix where each row corresponds to an individual test taker. The columns should represent the available features, including Age (ranging from 17 to 25), Gender (a binary variable: female or male), and Area (a binary variable: Wood Green or Burgess Hill). Additionally, the matrix includes a column for the response variable, indicating whether the individual passed their driving test (binary: 1 for passed, 0 for not passed).

```
log_model <- glm(Pass ~ Age + Area + Gender,
                 family = binomial, data = PassData)

summary(log_model)
```

```
##
## Call:
## glm(formula = Pass ~ Age + Area + Gender, family = binomial,
##      data = PassData)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.591480   0.043096   13.72  <2e-16 ***
## Age           -0.034700   0.002172  -15.97  <2e-16 ***
## AreaWood Green -0.259728   0.011119  -23.36  <2e-16 ***
## GenderMale      0.199755   0.010584   18.87  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 200720  on 145117  degrees of freedom
## Residual deviance: 199234  on 145114  degrees of freedom
## AIC: 199242
##
## Number of Fisher Scoring iterations: 4
```

```
predictions <- data.frame(
  Age = c(21, 21),
  Gender = c("Female", "Female"),
  Area = c("Burgess Hill", "Wood Green") # Keeping Age and Gender Constant
)

# Predicting probabilities
predicted_probs <- predict(log_model, newdata = predictions, type = "response")

# Add predictions to the data frame
predictions$Predicted_Probability <- predicted_probs

predictions
```

```
##   Age Gender      Area Predicted_Probability
## 1  21 Female Burgess Hill          0.4657467
## 2  21 Female   Wood Green          0.4020439
```

Finally, the expected passing rates for my friend at both test centres were calculated. With Age and Gender held constant, the Area variable was adjusted. The results showed a 6.4 percentage point advantage in favor of Burgess Hill, with the expected pass rate being **46.6% for Burgess Hill** and **40.2% for Wood Green**.

## Wald-test

We perform a Wald test on a term from a logistic regression model to assess the significance of the predictor variable “Area”

### Hypotheses for the Wald Test

The hypotheses being tested are as follows:

- Null Hypothesis ( $H_0$ ): The coefficient of **Area** is equal to 0, i.e., **Area** has no significant effect on the outcome.

$$H_0 : \beta_{\text{Area}} = 0$$

- Alternative Hypothesis ( $H_a$ ): The coefficient of **Area** is not equal to 0, i.e., **Area** has a significant effect on the outcome.

$$H_a : \beta_{\text{Area}} \neq 0$$

```
# Perform Wald Test
wald_result <- waldtest(log_model, terms = "Area")

# Print the result
print(wald_result)
```

```
## Wald test
##
## Model 1: Pass ~ Age + Area + Gender
## Model 2: Pass ~ Age + Gender
##   Res.Df Df      F    Pr(>F)
## 1 145114
## 2 145115 -1 545.69 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As a result, we reject the null hypothesis as there is significant evidence that the passing rates between the two sites.

## Conclusion and Limitations

Based on all my methods, **My suggestion is therefore: take the test in Burgess Hill** as the expected pass rate being **46.6% for Burgess Hill** and **40.2% for Wood Green**.

Now, I want to address some limitations of this analysis:

1. **Time-weighting:** While the data spans from **2008–09 to 2023–24**, I did not explicitly account for how long ago each test was taken. Tests from earlier years (e.g., 2008–09) may be less relevant compared to more recent data, yet equal weight was given to all individuals in the analysis. A potential improvement could involve **time-weighting**, where more recent data is given higher importance. Additionally, including predictors like the season or even weather conditions on the test day could enhance the model’s predictive power, provided such information is available.
2. **Model Assumptions:** The logistic regression, Wald test, and t-test rely on different assumptions and subsets of the data:
  - **Logistic Regression:**  
Logistic regression assumes that:

$$(Y_i | \text{Age}_i, \text{Area}_i, \text{Gender}_i) \sim \text{Bernoulli}(\theta_i)$$

It models all test takers within a single framework, incorporating predictors such as age, gender, and area.

- **Wald Test:**  
The Wald test examines whether the coefficient for the “Area” variable is significantly different from zero. It assumes that the outcome variable for each individual follows a Bernoulli distribution, based on the values of the predictors in the logistic regression model:

$$W_i | (\text{Area}, \text{other covariates}) \sim \text{Bernoulli}(\theta_W)$$

- **Two-Sample T-Test:**  
The t-test assumes that the pass rates for Burgess Hill and Wood Green are:
  - Normally distributed
  - Consist of independent observations.

Despite these differences, the three approaches complement each other:

- The logistic regression provides **concrete expected passing rates**.
- The Wald test confirms whether the **observed differences are statistically significant**.
- The t-test quantifies the **difference in means between groups**.

The fact that all methods arrived at the same conclusion ensures that the recommendation is robust and supported by multiple lines of evidence.

3. **Age Representation:** While age is treated as a continuous variable in the logistic regression model, the oldest group (25-year-olds) includes all individuals aged 25 and above. This aggregation likely includes a small number of much older individuals, who may have different passing rates, potentially inflating the Area coefficient in the model. Similarly, the results may be sensitive to other boundary conditions in the dataset, such as underrepresentation of very young or older individuals.
4. **Gender Bias:** Gender was included as a predictor in the logistic regression model, but potential gender biases in the data could still affect the results. For example:  
If certain testing areas systematically favor one gender (e.g., due to test design, societal norms, or examiner bias), this could skew the results.  
The model assumes that the effects of gender on passing rates are the same across areas, which might not hold true. Exploring interaction effects between gender and area could provide additional insights and help account for any such biases.