KINGSTON UNIVERSITY LONDON
DATA ANALYTICS AND VISUALIZATION
CI7330
COURSEWORK 2

NAME	KU NUMBER
Sujay Grama Suresh Kumar	K2201621

Module Leader: Dr. Rosie McNiece

Task 1

Summary of Insights:

- The analysis of the survey data reveals a diverse demographic landscape among different tenure categories. The age exhibits a uniform-distributed variation, with mean and median ages showcasing a broad distribution, spanning approximately 50 years across all categories.
- Income stability is evident, with consistent mean and median income values around £50,000, indicate a normal-distributed variation of financial well-being across all tenure categories. The Standard Deviation ranging approximately from 1.32 to 1.46, underscore the diverse income variations within each category.
- However, striking variations in education distribution are observed, particularly with private rent people having the highest percentage of individuals approximately 42% with the secondary education.
- The differences highlight the importance of customized campaign strategies for each category. The data-driven insights form a strong foundation for informed decision-making, helping leaders create effective and targeted political campaigns.

Tenure	Min Age	1st Quartile Age (25%)	Mean Age	Median Age	3rd Quartile Age (75%)	Max Age
Living in family member home	19	30	42.8	41	56	68
Owner	18	31	45.0	45	59	68
Private Rent	18	32	44.2	44	58	68
Social Rent	18	32	45.0	46	58.8	68

Table 1: Descriptive Statistics of Age by Tenure Category (Quantitative)

Tenure	Mean Income	Median Income	Min Income	Max Income	1st Quartile Income (25%)	3rd Quartile Income (75%)
Living in family member home	4.69	5	2	9	4	6
Owner	5.05	5	1	9	4	6
Private Rent	4.94	5	0	9	4	6
Social Rent	4.72	5	0	8	4	6

Table 2: Descriptive Statistics of Income by Tenure Category (Quantitative)

Tenure	Primary Education (%)	Secondary Education (%)	Higher Education (%)	Total (%)
Living in family member home	1.65% (33/2000)	2.15% (43/2000)	0.05% (1/2000)	3.85% (77/2000)
Owner	2.35% (47/2000)	22.7% (454/2000)	8.80% (176/2000)	33.85% (677/2000)
Private Rent	6.20% (124/2000)	30.95% (619/2000)	5.05% (101/2000)	42.20% (844/2000)
Social Rent	5.40% (108/2000)	13.75% (275/2000)	0.95% (19/2000)	20.10% (402/2000)

Table 3: Descriptive Statistics of Education by Tenure Category (Categorical)

Task 2

The clustered bar chart illustrates the distribution of education levels across different tenures. Primary education predominates in Owner, private rent, and social rent categories, while secondary education dominates in the Family tenure.

Code:

```
#Defining variable as a factor
dataset$tenure <- as.factor(dataset$tenure)
dataset$education <- as.factor(dataset$education)

legend_labels <- c("Primary", "Secondary", "Higher")

# Create a clustered bar chart
barplot(table(dataset$education, dataset$tenure),
      beside = TRUE,
      col = c("lightblue", "lightgreen", "lightcoral"),
      main = "Clustered Bar Chart of Tenure and Education",
      xlab = "Tenure",
      ylab = "Count")
legend("topleft", legend = legend_labels, fill = c("lightblue", "lightgreen", "lightcoral"), title =
"Education")
```

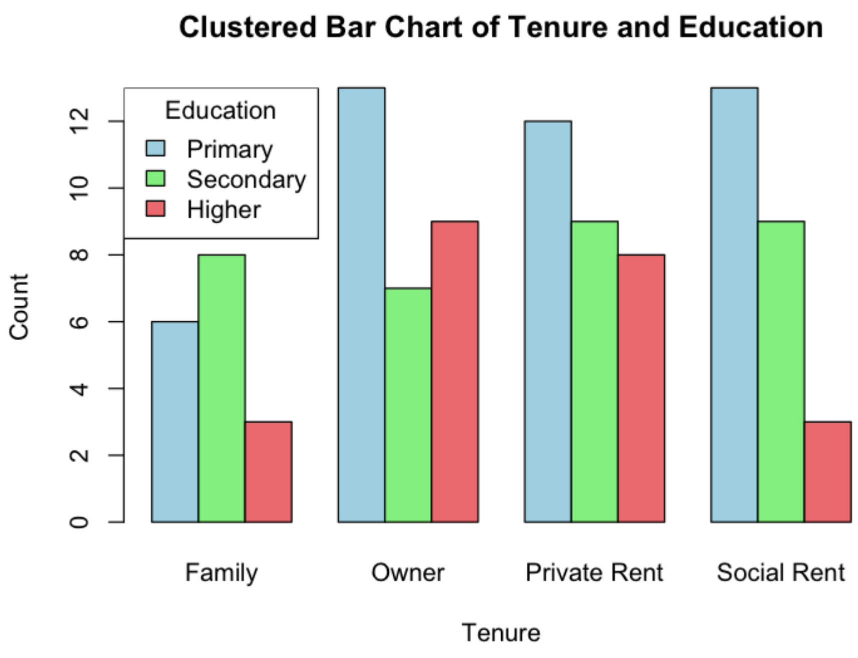

Figure 1: Clustered Bar Chart of Tenure and Education

Task 3

```
#Import Library
library(ggplot2)

# Load the dataset
dataset <- read.csv("dataset1_K2201621-1.csv")
head(dataset)

# Scatterplot with age on the x-axis and income on the y-axis
plot(x = dataset$age, y = dataset$income, xlab = "Age", ylab = "Income")

# Spline curves for each education category
primary_spline <- smooth.spline(x = dataset$age[dataset$education == 'Primary'], y =
dataset$income[dataset$education == 'Primary'], spar = 0.7)
secondary_spline <- smooth.spline(x = dataset$age[dataset$education == 'Secondary'], y =
dataset$income[dataset$education == 'Secondary'], spar = 0.7)
higher_spline <- smooth.spline(x = dataset$age[dataset$education == 'Higher'], y =
dataset$income[dataset$education == 'Higher'], spar = 0.7)

# Plot the spline curves
lines(primary_spline$x, primary_spline$y, col = 'purple', lwd = 2)
lines(secondary_spline$x, secondary_spline$y, col = 'orange', lwd = 2)
lines(higher_spline$x, higher_spline$y, col = 'green', lwd = 2)

# Save the plot as an image (adjust the filename and format as needed)
dev.copy(png, "scatterplot_with_splines.png")
dev.off()

#Add legend
legend("topleft", legend = c("Primary", "Secondary", "Higher"),
       col = c("purple", "orange", "green"), lty = 1, lwd = 2, cex = 0.8 )
```

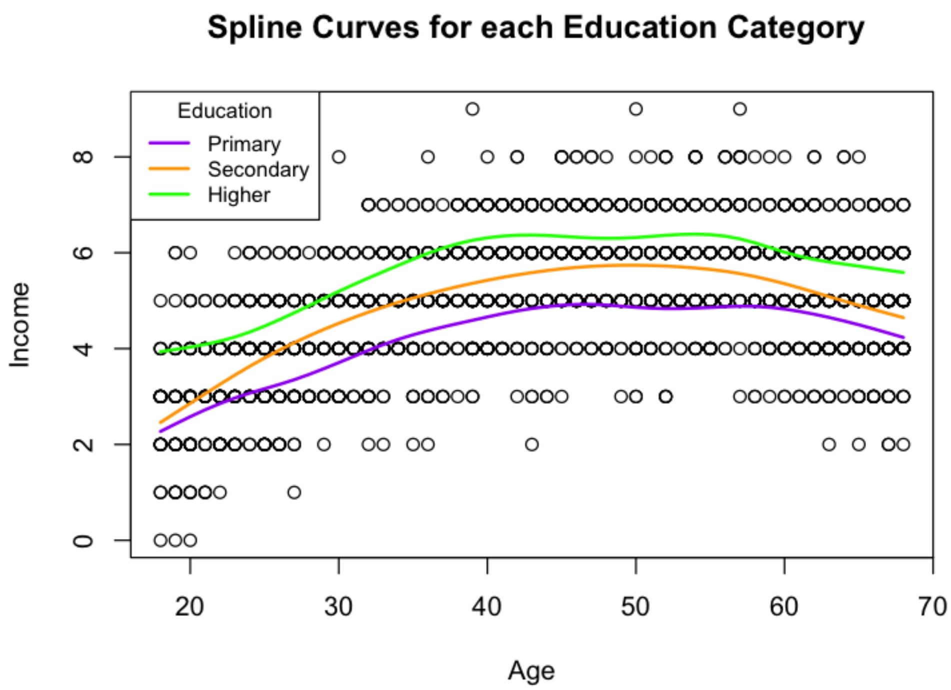

Figure 2: Spline Curves for each Education Category

Task 4

Linear Regression Model-1

In the present analysis, the logistic regression model employ ‘vote’ as the outcome variable, utilizing all predictor variables available in the dataset, namely ‘age’, ‘income’, ‘tenure’, ‘education’.

Formula:

glm(formula = vote ~ age + income + tenure + education, family = "binomial", data =dataset)

Summary of the Model

Coefficients:					
	Estimate	**Std. Error**	**Z value**	**P Value**	
(Intercept)	-3.0758242	0.2843344	-10.818	< 2e-16	***
age	-0.0002628	0.0042774	-0.061	0.95100	
income	0.2783209	0.049153	5.662	1.49e-08	***
tenureFam	0.1827145	0.3319767	0.550	0.58206	
tenurePri	0.3562232	0.1340113	2.658	0.00786	**
tenureSoc	0.0843532	0.1761333	0.479	0.63200	
educationPrimary	-0.1560159	0.1838496	-0.849	0.39610	
educationHigher	0.7138452	0.1507457	4.735	2.19e-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Table 4: Summary of Linear Regression Model-1

AIC: 1931.6
Residual deviance: 1915.6 on 1992 degrees of freedom
Null deviance: 2007.1 on 1999 degrees of freedom

The summary of the Logistic Regression Model-1 indicates that income, education, and tenure variables are statistically significant predictors of the voting outcome ('vote'), with p-values less than 0.05. Conversely, age appears to be an insignificant predictor. To prevent overfitting the 'age' is excluded from the model.

Linear Regression Model-2

In the subsequent analysis, the logistic regression model focuses on the significant predictor variables identified in Model-1. The outcome variable 'vote' is considered with the predictors 'Income,' 'Tenure,' and 'Education.'

Formula:

glm(formula = vote ~ income + tenure + education, family = "binomial", data =dataset)

Summary of the Model

Coefficients:					
	Estimate	**Std. Error**	**Z value**	**P value**	
(Intercept)	-3.08242	0.26345	-11.700	< 2e-16	***
income	0.27724	0.04592	6.038	1.56e-09	***
tenureFam	0.18315	0.33191	0.552	0.58108	
tenurePri	0.35651	0.13393	2.662	0.00777	**
tenureSoc	0.08420	0.17611	0.478	0.63260	
educationPrimary	-0.15641	0.18374	-0.851	0.39462	
educationHigher	0.71453	0.15034	4.753	2.01e-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Table 5: Summary of Linear Regression Model-2

AIC: 1929.6
Residual deviance: 1915.6 on 1993 degrees of freedom
Null deviance: 2007.1 on 1999 degrees of freedom

The comparison of Linear Regression Model-1 and Linear Regression Model-2 based on the Akaike Information Criterion (AIC) suggests that Model-2 performs slightly better than Model-1 as it exhibits a lower AIC value. This implies that Model-2 provides a more straightforward representation of the data, balancing goodness of fit with model simplicity. Therefore, Model-2 is preferred over Model-1 for its improved model fit. Henceforth, best Linear Regression Model-2 can be considered as the primary model.

Confidence Interval for the predictor variables

Code Snippet**:**

```
confidence_intervals <- confint(model_1)
confidence_intervals
```

Predictor	**conf_int. 2.5%**	**conf_int. 97.5%**
(Intercept)	-3.6069971	-2.5738388

income	0.1880156	0.3680932
tenureFam	-0.5047946	0.8058015
tenurePri	0.0955076	0.6208699
tenureSoc	-0.2642701	0.4270003
educationPrimary	-0.5263990	0.1953839
educationHigher	0.4181612	1.0079510

Table 6: Confidence Interval

The confidence intervals for all the predictors in Linear Regression Model-2 give a range of likely effects. For instance, the starting point (intercept) for the prediction could be anywhere from -3.61 to -2.57.

When income goes up by one unit, the prediction might change between 0.19 and 0.37. While family and social tenure may not have a clear impact (their intervals include zero), private tenure seems to matter, with an effect ranging from 0.10 to 0.62.

Primary education might not be significant, with an effect from -0.53 to 0.20, but higher education is noteworthy, showing an impact from 0.42 to 1.01. The provided intervals aid in comprehending the level of uncertainty or certainty associated with the influence of each factor on predictions.

Marginal Effect and Confidence Intervals using 'marginaleffects' package

Code Snippet**:**

```
#Import necessary libraries
library(marginaleffects)

marginal_effects <- marginaleffects(model_1)
summary(marginal_effects)
```

factor	**AME**	**SE**	**z**	**p**	**lower**	**upper**
educationHigher	0.1269	0.0295	4.2952	0.0000	0.0690	0.1848
educationPrimary	-0.0218	0.0248	-0.8805	0.3786	-0.0704	0.0268
income	0.0424	0.0069	6.1538	0.0000	0.0289	0.0559
tenureFam	0.0267	0.0504	0.5299	0.5962	-0.0721	0.1255
tenurePri	0.0546	0.0201	2.7113	0.0067	0.0151	0.0940
tenureSoc	0.0119	0.0251	0.4750	0.6348	-0.0373	0.0611

Table 7: Marginal Effects

factor	***Factor variable***
AME	***Average Marginal Effect***
SE	***Standard Error***
z	***z-value***
p	***p-value***
lower	***Lower bound of Confidence Interval (2.5%)***
upper	***Upper bound of Confidence Interval (97.5%)***

Table 8: Abbreviations for Marginal Effect (Table 7)

- Education:
  - Higher education significantly boosts the probability of a favourable outcome by 12.69% as p-value is less than 0.05 ($p<0.05$). The 95% confidence interval (CI) ranges from 6.90% to 18.48%, while primary education, though not statistically significant, suggests a marginal decrease of 2.18% in the likelihood of a positive outcome (CI: -7.04% - 2.68%).
  - Hence, higher levels of education and income significantly enhance the likelihood of positive outcomes, providing individuals with increased opportunities and resources for favourable life circumstances.
- Income:
  - Each unit increase in income substantially raises the probability of a positive outcome by 4.24% (CI: 2.89% - 5.59%).
- Tenure:
  - Private tenure significantly enhances the likelihood of a positive outcome by 5.46% ($p<0.05$).The 95% CI ranges from 1.51% to 9.40%, while family tenure shows a slight non-significant increase of 2.67% in the probability of a positive outcome (CI: -7.21% - 12.55%). Social tenure indicates a marginal and non-significant increase of 1.19% (CI: -3.73% - 6.11%).

These consolidated view helps for the party leadership, emphasizing the impactful role of education, income, and tenure on voter preferences, facilitating strategic campaign planning.

Task 5

The general logistic regression predictive formula in terms of log odds is:

$$\log w_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

The formula for estimating the log-odds of the intention to vote for the Data-Driven Party is:

$$\text{Log-Odds} = \beta_0 + \beta_1 \times \text{Income} + \beta_2 \times \text{Tenure} + \beta_3 \times \text{Education}$$

Where :

β_0 is the intercept or constant term.

β_1 ,β_1, β_3 are the coefficients associated with the predictor variables Income, Tenure, Education.

#Predict log-odds function applying to dataset2
log_odds <- predict(model_1, newdata = dataset2, type = "response")
This function gives the predicted probabilities for happy town.
# Convert log-odds to probabilities
predicted_probabilities <- plogis(log_odds)

The plogis function is then used to convert the log-odds to probabilities. Now, predicted_probabilities contains the predicted probabilities for "Happytown"

Probability Comparison Table for the dataset2 (example):

Vote	Tenure	Education	Age	Income	Predprob	New_pred_prob	Difference
0	Pri	Secondary	29	4	0.18090846	0.165628354	-0.0152801
1	Own	Secondary	56	6	0.20458692	0.194824303	-0.0097626
0	Fam	Primary	64	4	0.10359505	0.124912739	0.02131769

Table 8: Predicted Probability Comparison (dataset2)

In summary, when the given predicted probabilities are compared with the newly calculated predicted probabilities, there are very slight differences between the original values and the ones that have been newly computed. These variations are extremely small, so they can be considered negligible. The overall consistency between the two sets of predictions highlights the strength and dependability of the model in estimating the likelihood of people voting for the Data-Driven Party in Happytown. Therefore, the model's reliability and suitability for the given application are affirmed.

Codes for all the tasks:

Task 1:

```
# Load necessary libraries
library(dplyr)

# Read the dataset
dataset <- read.csv("dataset1_K2201621-1.csv")

# Task 1: Descriptive statistics for age, income, and education for each category of tenure
tenure_groups <- dataset %>%
  group_by(tenure)

# Descriptive statistics for age
age_stats <- tenure_groups %>%
  summarize(
    mean_age = mean(age),
    median_age = median(age),
    sd_age = sd(age),
    min_age = min(age),
    max_age = max(age),
    range_age = max(age) - min(age),
    Q1_age = quantile(age, 0.25),
    Q3_age = quantile(age, 0.75),
  )

# Descriptive statistics for income
income_stats <- tenure_groups %>%
  summarize(
    mean_in = mean(income),
    median_in = median(income),
    sd_in = sd(income),
    min_in = min(income),
    max_in = max(income),
    range_age = max(income) - min(income),
    Q1_In = quantile(income, 0.25),
    Q3_In = quantile(income, 0.75),
  )

# Descriptive statistics for education
descriptive_stats <- dataset %>%
  group_by(tenure) %>%
  summarise(
    Primary_count = sum(education == 'Primary'),
    Primary_percent = sum(education == 'Primary') / 2000,

    Secondary_count = sum(education == 'Secondary'),
    Secondary_percent = sum(education == 'Secondary') / 2000,

    Higher_count = sum(education == 'Higher'),
    Higher_percent = sum(education == 'Higher') / 2000,
```

```
    total_count = n(),
    total_percent = n()/2000
  )

# Display the results
print("Frequency, Cumulative Frequency, Percent, Cumulative Percent for each category of education with respect to tenure")
print(education_stats)

# Display the results
cat("Task 1: Descriptive statistics for age, income, and education for each category of tenure\n\n")
print(age_stats)
print(income_stats)
print (descriptive_stats)
```

Task 2:

```
#Defining variable as a factor
dataset$tenure <- as.factor(dataset$tenure)
dataset$education <- as.factor(dataset$education)

legend_labels <- c("Primary", "Secondary", "Higher")
# Create a clustered bar chart
barplot(table(dataset$education, dataset$tenure),
        beside = TRUE,
        col = c("lightblue", "lightgreen", "lightcoral"),
        main = "Clustered Bar Chart of Tenure and Education",
        xlab = "Tenure",
        ylab = "Count")
legend("topleft", legend = legend_labels, fill = c("lightblue", "lightgreen", "lightcoral"), title = "Education")
```

Task 3:

```
#Import Library
library(ggplot2)

# Load the dataset
dataset <- read.csv("dataset1_K2201621-1.csv")
head(dataset)

# Scatterplot with age on the x-axis and income on the y-axis
plot(x = dataset$age, y = dataset$income, xlab = "Age", ylab = "Income")

# Spline curves for each education category
primary_spline <- smooth.spline(x = dataset$age[dataset$education == 'Primary'], y = dataset$income[dataset$education == 'Primary'], spar = 0.7)
secondary_spline <- smooth.spline(x = dataset$age[dataset$education == 'Secondary'], y = dataset$income[dataset$education == 'Secondary'], spar = 0.7)
higher_spline <- smooth.spline(x = dataset$age[dataset$education == 'Higher'], y = dataset$income[dataset$education == 'Higher'], spar = 0.7)
# Plot the spline curves

lines(primary_spline$x, primary_spline$y, col = 'purple', lwd = 2)
```

```
lines(secondary_spline$x, secondary_spline$y, col = 'orange', lwd = 2)
lines(higher_spline$x, higher_spline$y, col = 'green', lwd = 2)
# Save the plot as an image (adjust the filename and format as needed)
dev.copy(png, "scatterplot_with_splines.png")
dev.off()

#Add legend
legend("topleft", legend = c("Primary", "Secondary", "Higher"),
      col = c("purple", "orange", "green"), lty = 1, lwd = 2, cex = 0.8 )
```

Task 4:

```
# Load the required libraries
library(ggplot2)
library(MASS)  # Required for the 'glm' function
library(marginaleffects)
library(margins)

# Convert categorical variables to factors
data$tenure <- factor(data$tenure, levels = c("Own", "Fam", "Pri", "Soc"))
data$education <- factor(data$education, levels = c("Secondary", "Primary", "Higher"))

# Fitting logistic regression model-1
model <- glm(vote ~ age + income + tenure + education, data = data, family = "binomial")
# Show coefficients and confidence intervals
summary(model)

# Fit new logistic regression model-2 by dropping 'age' (p>0.05)
model_1 <- glm(vote ~ income+ tenure + education, data = data, family = "binomial")
summary(model_1)

#Confidence Interval for the model
conf_intervals <- confint(model_1)
print(conf_intervals)

#Calculate marginal effects
marginal_effects <- marginaleffects(model_1)
summary(marginal_effects)
```

Task 5:

```
# Read the dataset2
dataset2 <- read.csv("dataset2_K2201621-3.csv")
# Predict log-odds for dataset2
log_odds <- predict(model_1, newdata = dataset2, type = "response")
dataset2$new_pred_prob <- log_odds
dataset2$difference <- dataset2$new_pred_prob – dataset2$predprob
# Transform log-odds to probabilities
predicted_probabilities <- plogis(log_odds)
predicted_probabilities

# Display the results
result <- data.frame(LogOdds = log_odds, Probability = predicted_probabilities)
print(result)
```