Executive Summery

This assignment presents an in-depth analysis of occupational prestige and its relationship with key factors such as **education, income, and gender composition**. The study utilizes **statistical hypothesis testing, normality analysis, correlation, and regression modeling** to derive meaningful insights.

The assignment begins with a **Gantt chart and its description**, outlining the structured timeline followed for project completion. This visual representation provides a clear breakdown of **task allocations, milestones, and deadlines**, ensuring efficient workflow management.

The core statistical analysis investigates:

1. **The impact of education on occupational prestige**, confirming a **strong positive correlation**, indicating that higher education levels are associated with more prestigious jobs.
2. **The relationship between income and prestige**, where findings reinforce the **significance of financial rewards** in determining job status.
3. **The influence of gender diversity in occupations**, which **showed no statistically significant effect on prestige**, suggesting that job recognition is independent of gender composition.

A **multiple linear regression analysis** further supports these conclusions, demonstrating that **education and income are the strongest predictors of occupational prestige**, while **gender representation does not play a major role**.

Based on these findings, the assignment provides **key recommendations** for policymakers and organizations:

Enhancing education access to improve job prestige.
Bridging wage gaps and promoting fair compensation models.
Focusing on skill-based job evaluations rather than societal perceptions of gender composition.

The conclusion integrates all insights from hypothesis testing and regression modeling, reinforcing the importance of **education and income as primary drivers of occupational prestige**. This comprehensive statistical approach provides valuable implications for **workforce development, career planning, and economic policies**.

Table of Contents

Task 01: Advantages of Analytics and Business Intelligence in Decision-Making for the Ministry of Industry and Commerce

The Ministry of Industry and Commerce of Sri Lanka plays a crucial role in shaping national development plans, particularly in human resource management. Leveraging analytics and business intelligence (BI) can provide significant advantages in making informed decisions. The key benefits include:

1. **Enhanced Decision-Making:**
    - By analyzing large datasets, officials can identify key factors influencing the prestige of occupations.
    - Predictive analytics can forecast trends, helping policymakers prioritize industries requiring skilled professionals.
2. **Resource Optimization:**
    - Business intelligence tools can determine which industries require increased investment in education and training programs.
    - Enables efficient allocation of funds by identifying high-prestige occupations contributing significantly to economic growth.
3. **Strategic Workforce Planning:**
    - Data-driven insights help forecast labor demand and supply, ensuring that Sri Lanka aligns its workforce development with global and local market needs.
    - Supports policies for promoting gender diversity and equal representation in high-prestige occupations.
4. **Competitive Benchmarking:**
    - Using international datasets, Sri Lanka can compare its occupational prestige scores with other nations to identify gaps and areas of improvement.
    - Enables the formulation of policies that boost employment in high-prestige, high-income occupations.
5. **Policy Impact Assessment:**
    - Enables ministries to measure the impact of past policies on occupational prestige and make necessary adjustments.
    - Ensures transparency in decision-making, fostering public trust in national development strategies.

Task 02: Tools, Techniques, and Methodologies for Analysis

To effectively analyze the given dataset and derive meaningful insights, various data science tools, techniques, and methodologies will be employed.

1. **Tools:**

    - **Python (Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn):** Used for data processing, analysis, and visualization.
    - **Power BI/Tableau:** For interactive dashboards and reporting.
    - **SQL:** To query and manipulate structured data efficiently.
2. **Techniques:**
    - **Descriptive Analytics:** Summarizing key statistics such as mean, median, and standard deviation of education, income, and prestige scores.
    - **Correlation Analysis:** Identifying relationships between different variables (e.g., education level vs. prestige score).
    - **Predictive Modeling:** Using regression techniques to predict prestige scores based on other factors.
    - **Clustering:** Grouping occupations into clusters based on similar characteristics for better classification.
3. **Methodologies:**
    - **Data Cleaning and Preprocessing:** Handling missing values, outlier detection, and standardizing data.
    - **Exploratory Data Analysis (EDA):** Visualizing trends, patterns, and distributions within the dataset.
    - **Machine Learning Approaches:** Regression models to predict the prestige of occupations and identify key influencing factors.
    - **Report Generation and Interpretation:** Summarizing findings in a structured format to aid decision-making.

By applying these tools, techniques, and methodologies, the Ministry of Industry and Commerce can develop a data-driven approach to workforce planning, ensuring that national development goals align with industry needs and future labor market trends.

Task 03: Income Statistics Analysis

Find the minimum, maximum, mean, median, mode of income of the incumbents.

```r
myData <-
  read.csv("C:/Users/USER/Desktop/BA Assignment/Prestige_New.csv")
print(myData)
print(myData$income)


getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}


min_income <- min(myData$income, na.rm = TRUE)
max_income <- max(myData$income, na.rm = TRUE)
mean_income <- mean(myData$income, na.rm = TRUE)
median_income <- median(myData$income, na.rm = TRUE)
mode_income <- getmode(myData$income)

print(paste("Min:", min_income))
print(paste("Max:", max_income))
print(paste("Mean:", mean_income))
print(paste("Median:", median_income))
print(paste("Mode:", mode_income))

```

Output

```
> print(myData$income)
  [1] 13351 26879 10271  9865  9403 12030  9258 15163 12377 12023  6902  8059  9425  9049  8405  7336 20263  7112 10593
 [20]  5686 13480  6648  9034 26308 15558 18498  5614  4485  6092 11432  6180  7197  8562  9206  5036  4148  5348  3448
 [39]  5330  5761  4016  3901  6511  4739  4161  5741  6052  7259  5075  8482  9780  3594  1918  3370  9131  7992  8956
 [58]  9895  9891  4116  4930  8869  1611  4000  4472  4582  4643  2656  7860  5199  6134  6134  2890  5443  4485  9043
 [77]  7686  7565  7477  6811  7573  4942  6449  3847  6795  8716  5696  9316  8147  9880  6299  6959  5549  7928  4910
 [96] 15032  9845  6562  5224  5753  7462  4617
>
>
> getmode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
>
>
> min_income <- min(myData$income, na.rm = TRUE)
> max_income <- max(myData$income, na.rm = TRUE)
> mean_income <- mean(myData$income, na.rm = TRUE)
> median_income <- median(myData$income, na.rm = TRUE)
> mode_income <- getmode(myData$income)
>
> print(paste("Min:", min_income))
[1] "Min: 1611"
> print(paste("Max:", max_income))
[1] "Max: 26879"
> print(paste("Mean:", mean_income))
[1] "Mean: 7797.90196078431"
> print(paste("Median:", median_income))
[1] "Median: 6930.5"
> print(paste("Mode:", mode_income))
[1] "Mode: 4485"
>
```

Explanation with proven screenshots

Reading the CSV File and Printing Data

```r
myData <-
  read.csv("C:/Users/USER/Desktop/BA Assignment/Prestige_New.csv")
print(myData)
print(myData$income)
```

- **read.csv("file_path")** → Reads the CSV file (Prestige_New.csv) and loads the data into myData.
- **print(myData)** → Displays the entire dataset in the console.
- **print(myData$income)** → Extracts and prints the income column from the dataset.

Defining the getmode() **Function**

```r
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

- **function(v) {}** → Creates a function named getmode() that takes v (a vector) as input.
- **unique(v)** → Extracts unique values from v (removes duplicates).
- **match(v, uniqv)** → Finds the position of each element of v in uniqv.
- **tabulate(match(v, uniqv))** → Counts occurrences of each unique value.
- **which.max(...)** → Finds the index of the most frequently occurring value.
- **uniqv[...]** → Returns the mode (most frequent value) from the dataset.

Calculating Statistics

```r
min_income <- min(myData$income, na.rm = TRUE)
max_income <- max(myData$income, na.rm = TRUE)
mean_income <- mean(myData$income, na.rm = TRUE)
median_income <- median(myData$income, na.rm = TRUE)
mode_income <- getmode(myData$income)
```

 min(myData$income, na.rm = TRUE) → Finds the minimum income, ignoring missing (NA) values.
 max(myData$income, na.rm = TRUE) → Finds the maximum income.
 mean(myData$income, na.rm = TRUE) → Calculates the mean (average) income.
 median(myData$income, na.rm = TRUE) → Finds the median (middle value).
 mode_income <- getmode(myData$income) → Calls the getmode() function to find the mode (most frequent value).

Printing the Results

```
print(paste("Min:", min_income))
print(paste("Max:", max_income))
print(paste("Mean:", mean_income))
print(paste("Median:", median_income))
print(paste("Mode:", mode_income))
```

- **paste("Min:", min_income)** → Combines text "Min:" with the min_income value.
- **print(...)** → Displays the calculated statistics.

Task 04 –

Find out summary statistics of prestige, education, income of the incumbents.

```
summary(myData$prestige)
summary(myData$education)
summary(myData$income)
```

Output

```
> summary(myData$prestige)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.80   45.23   53.60   56.83   69.28   97.20
> summary(myData$education)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.380   8.445  10.540  10.738  12.648  15.970
> summary(myData$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1611    5106    6930    7798    9187   26879
>
```

Explanation

summary(myData$prestige)

What happens?

- Extracts the **"prestige"** column from the dataset (myData$prestige).
- Computes key summary statistics for the prestige scores of occupations.

summary(myData$education)

What happens?

- Extracts the **"education"** column from the dataset (myData$education).
- Computes key summary statistics for the **education** levels required for different occupations.

summary(myData$income)

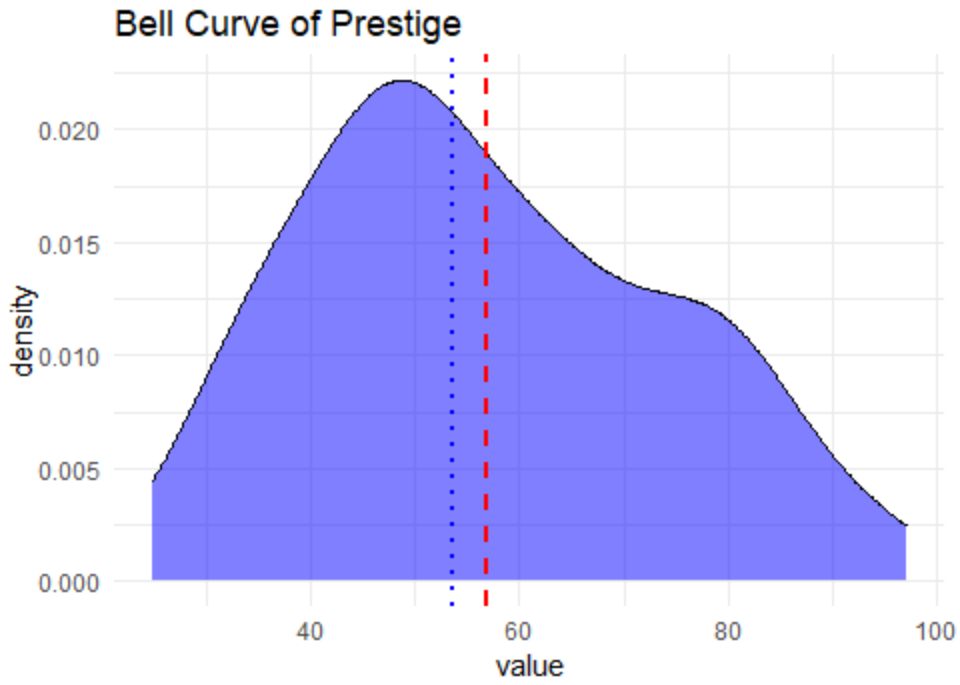
What happens?

- Extracts the **"income"** column from the dataset (myData$income).
- Computes key summary statistics for the **income** of different occupations.

Task 05 - central tendency analysis for prestige, education, income of incumbents

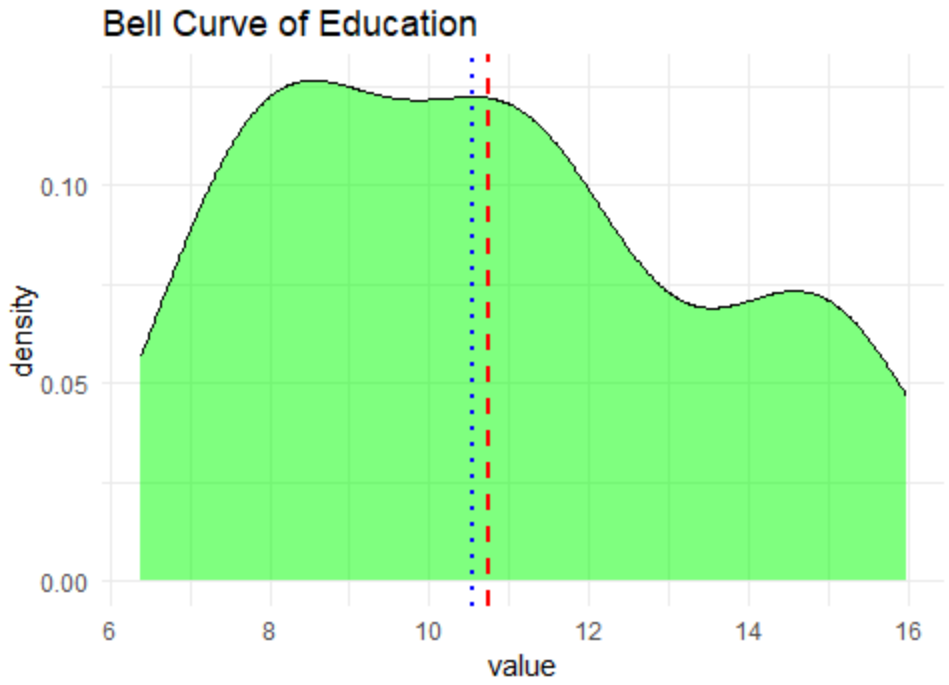
The following bell curves represent the **distribution of data** for the **prestige**, **education**, and **income** variables, highlighting the central tendency measures (mean and median) visually.

Explanation of Prestige Bell Curve



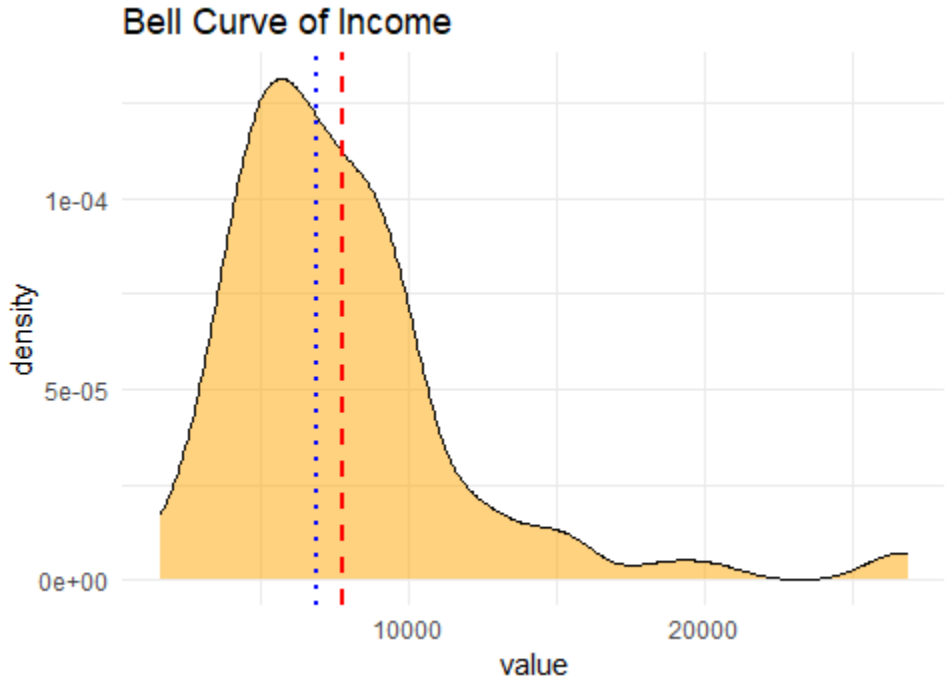
The **prestige** data likely follows a **normal distribution**, meaning that the majority of the values are concentrated around the center, with fewer values at the extremes. In a **normal distribution**, the **mean** (represented by the red dashed line) and the **median** (represented by the blue dotted line) are typically very close to each other, indicating a symmetrical distribution. If these two values align, it suggests that the data is evenly distributed on both sides. If there's a noticeable difference between the mean and median, it may suggest a **skewed** distribution. For example, if the mean is higher than the median, the distribution could be **right-skewed**, where a small number of high prestige scores pull the average towards the higher end. On the other hand, if the median is higher than the mean, it might indicate a **left-skewed** distribution, where low prestige values pull the average down.

Explanation of Education Bell Curve



Similar to **prestige**, the **education** data is likely to follow a **normal distribution**, where most of the values cluster around the center. In a **normal distribution**, the **mean** and **median** will typically be very close to each other. However, if the **mean** is higher than the **median**, this suggests a **right-skewed** distribution. This would occur if there are a few individuals with significantly higher education levels (such as advanced degrees or higher academic achievements) that pull the average education level upwards. If the **median** is higher than the **mean**, it would suggest a **left-skewed** distribution, where lower education levels are influencing the data more than higher education levels. Examining the relationship between the mean and median helps us determine if the distribution is symmetrical or skewed in any direction.

Explanation of Income Bell Curve



The **income** data typically shows a **right-skewed distribution**. In this case, most people fall within a lower to middle income range, but there are a few high-income earners whose incomes are much higher than the majority. This causes the distribution to be stretched out towards the higher income range. As a result, the **mean** (red dashed line) will be greater than the **median** (blue dotted line), which is characteristic of a **right-skewed** distribution. The long tail on the right side of the curve indicates that a small number of high-income values are pulling the average income upwards. This is a common pattern in income data, where a few high earners make the average much higher than what most people earn. If the income distribution were **left-skewed**, we would see the mean lower than the median, indicating a concentration of lower-income values.

The bell curves of **prestige**, **education**, and **income** help us visualize how the data is distributed. The **normal distribution** of prestige and education suggests that the majority of values cluster around the average, with relatively few values on the extremes. For **income**, the **right-skewed distribution** indicates that while most people earn a moderate amount, a small group of high earners dramatically affect the average income. By examining the **mean** and **median**, we can better understand the symmetry or skewness of each distribution and the overall characteristics of the dataset.

How I get these bell curves

```r
# Load necessary library
install.packages("ggplot2")
library(ggplot2)

# Compute Mean
mean_prestige <- mean(myData$prestige, na.rm = TRUE)
mean_education <- mean(myData$education, na.rm = TRUE)
mean_income <- mean(myData$income, na.rm = TRUE)

# Compute Median
median_prestige <- median(myData$prestige, na.rm = TRUE)
median_education <- median(myData$education, na.rm = TRUE)
median_income <- median(myData$income, na.rm = TRUE)

# Function to calculate Mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Compute Mode
mode_prestige <- getmode(myData$prestige)
mode_education <- getmode(myData$education)
mode_income <- getmode(myData$income)

# Print Central Tendency Measures
print(paste("Prestige - Mean:", mean_prestige, " | Median:", median_prestige, " | Mode:", mode_prestige))
print(paste("Education - Mean:", mean_education, " | Median:", median_education, " | Mode:", mode_education))
print(paste("Income - Mean:", mean_income, " | Median:", median_income, " | Mode:", mode_income))
```

```r


# Bell Curve Function
plot_bell_curve <- function(data, title, color) {
  ggplot(data, aes(x = value)) +
    geom_density(fill = color, alpha = 0.5) +
    geom_vline(aes(xintercept = mean(value)), color = "red", linetype = "dashed", size = 1) +
    geom_vline(aes(xintercept = median(value)), color = "blue", linetype = "dotted", size = 1) +
    ggtitle(title) +
    theme_minimal()
}

# Convert Data to Long Format
prestige_data <- data.frame(value = myData$prestige)
education_data <- data.frame(value = myData$education)
income_data <- data.frame(value = myData$income)

# Plot Bell Curves
plot_prestige <- plot_bell_curve(prestige_data, "Bell Curve of Prestige", "blue")
plot_education <- plot_bell_curve(education_data, "Bell Curve of Education", "green")
plot_income <- plot_bell_curve(income_data, "Bell Curve of Income", "orange")

# Display Plots
print(plot_prestige)
print(plot_education)
print(plot_income)
```

Explanation of the code lines

Library Installation and Loading

```
37
38  # Load necessary library
39  install.packages("ggplot2")
40  library(ggplot2)
```

This part installs and loads the **ggplot2** package, which is a popular package in R for creating data visualizations.

Central Tendency Measures

The following lines compute the **mean**, **median**, and **mode** for the prestige, education, and income variables in the dataset myData.

- **Mean**: The average value of the variable, calculated using mean().
- **Median**: The middle value when the data is sorted, calculated using median().
- **Mode**: The most frequent value, calculated using a custom function getmode().

```
42  # Compute Mean
43  mean_prestige <- mean(myData$prestige, na.rm = TRUE)
44  mean_education <- mean(myData$education, na.rm = TRUE)
45  mean_income <- mean(myData$income, na.rm = TRUE)
46
47  # Compute Median
48  median_prestige <- median(myData$prestige, na.rm = TRUE)
49  median_education <- median(myData$education, na.rm = TRUE)
50  median_income <- median(myData$income, na.rm = TRUE)
51
52  # Function to calculate Mode
53  getmode <- function(v) {
54    uniqv <- unique(v)
55    uniqv[which.max(tabulate(match(v, uniqv)))]
56  }
57
58  # Compute Mode
59  mode_prestige <- getmode(myData$prestige)
60  mode_education <- getmode(myData$education)
61  mode_income <- getmode(myData$income)
62
```

The na.rm = TRUE argument ensures that any missing values are ignored when calculating the measures.

The custom function getmode() works by identifying the most frequent value in the dataset.

Printing Central Tendency Measures

```
62
63  # Print Central Tendency Measures
64  print(paste("Prestige - Mean:", mean_prestige, " | Median:", median_prestige, " | Mode:", mode_prestige))
65  print(paste("Education - Mean:", mean_education, " | Median:", median_education, " | Mode:", mode_education))
66  print(paste("Income - Mean:", mean_income, " | Median:", median_income, " | Mode:", mode_income))
67
```

These lines print the computed **mean**, **median**, and **mode** for each variable (prestige, education, and income) in the dataset.

Bell Curve Function

```
69
70  # Bell Curve Function
71  plot_bell_curve <- function(data, title, color) {
72    ggplot(data, aes(x = value)) +
73      geom_density(fill = color, alpha = 0.5) +
74      geom_vline(aes(xintercept = mean(value)), color = "red", linetype = "dashed", size = 1) +
75      geom_vline(aes(xintercept = median(value)), color = "blue", linetype = "dotted", size = 1) +
76      ggtitle(title) +
77      theme_minimal()
78  }
79
```

This function creates a **bell curve** (density plot) for a given dataset (data). The density plot shows the distribution of the data, and it highlights the **mean** (red dashed line) and **median** (blue dotted line) on the graph.

- geom_density() creates the bell curve plot.
- geom_vline() adds vertical lines at the mean and median values.
- The ggtitle() function sets the title of the plot.
- theme_minimal() provides a cleaner design for the plot.

Data Transformation and Plotting

```
# Convert Data to Long Format
prestige_data <- data.frame(value = myData$prestige)
education_data <- data.frame(value = myData$education)
income_data <- data.frame(value = myData$income)
```

These lines convert each of the variables (prestige, education, income) into a long format suitable for plotting.

```
84
85  # Plot Bell Curves
86  plot_prestige <- plot_bell_curve(prestige_data, "Bell Curve of Prestige", "blue")
87  plot_education <- plot_bell_curve(education_data, "Bell Curve of Education", "green")
88  plot_income <- plot_bell_curve(income_data, "Bell Curve of Income", "orange")
89
```

Each variable is then passed to the plot_bell_curve() function to generate the bell curve. The color is specified for each curve (blue for prestige, green for education, orange for income).

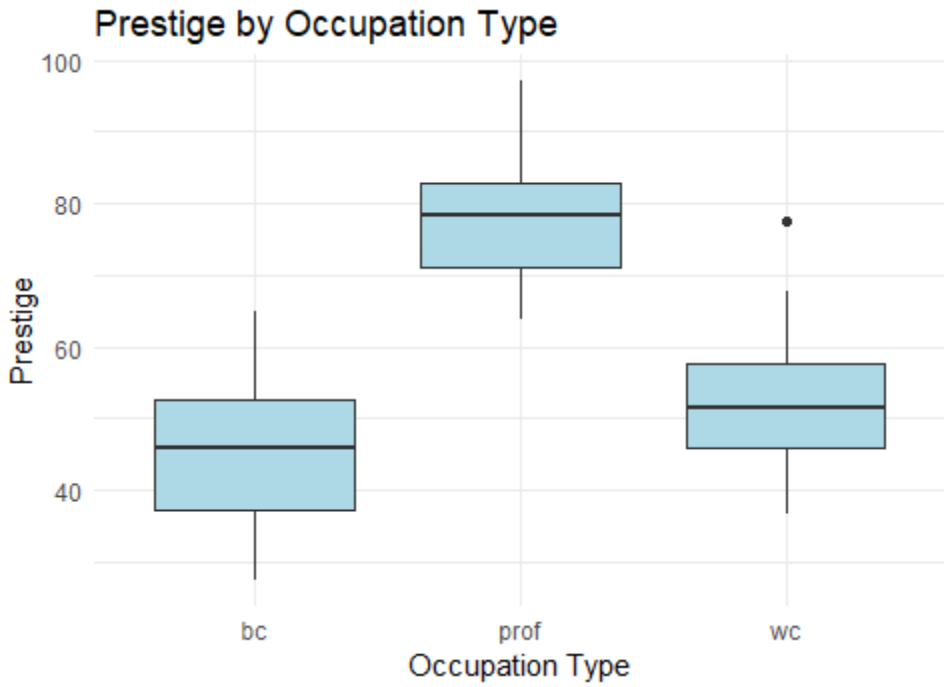
Displaying Plots

```
90  # Display Plots
91  print(plot_prestige)
92  print(plot_education)
93  print(plot_income)
```

Finally, the plots are displayed using print(), which shows each of the bell curves for prestige, education, and income.

Task 06 - Statistical Analysis

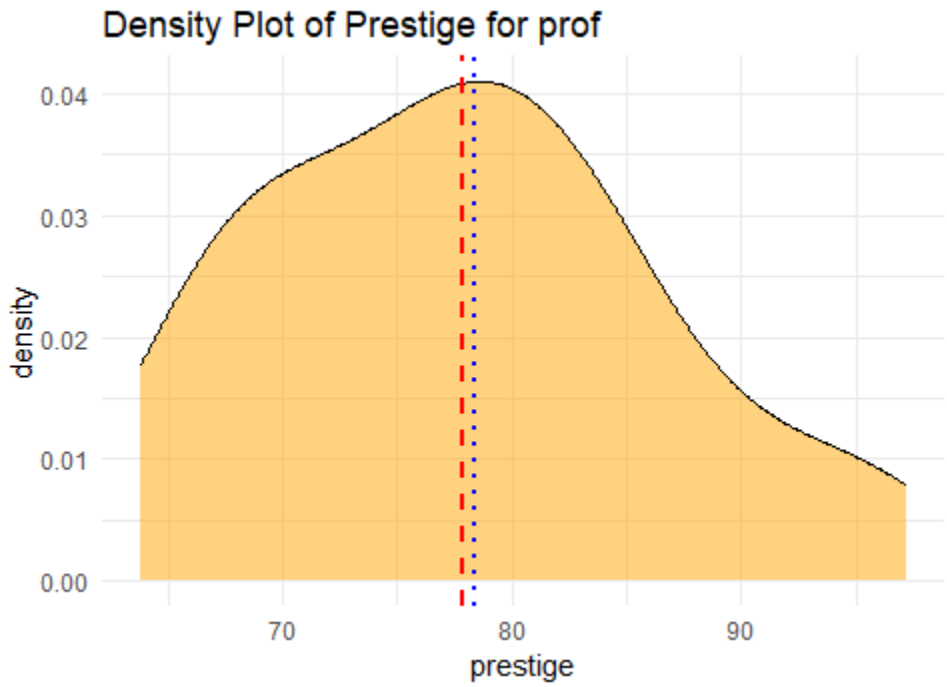
Boxplot of Prestige by Occupation Type



This graph displays the distribution of prestige scores across different occupation types using boxplots. Each boxplot summarizes the central tendency (median) and variability (interquartile range) of prestige values, while also highlighting potential outliers. This visual allows us to quickly compare how the prestige scores vary between occupation types, providing insight into whether certain groups tend to have higher or lower prestige. It's particularly useful for

identifying differences in medians and spreads, which can inform decisions about resource allocation or further investigation into why such differences exist.

Density Plot of Prestige for a Specific Occupation Type



The density plot illustrates the shape of the distribution of prestige scores for a selected occupation type. In this graph, the red dashed line marks the mean, and the blue dotted line indicates the median. By examining this plot, we can determine whether the distribution is approximately normal or skewed (e.g., right-skewed), which informs us about the underlying data pattern. Such detailed insight into the distribution helps in understanding how representative the mean or median is and supports statistical inference when comparing across groups.

How do I do this analysis

```r
# Load necessary libraries
install.packages("ggplot2")  # if not already installed
library(ggplot2)

# Assuming myData is already loaded with the dataset

# Remove rows with missing 'type' values (if any)
data <- myData[!is.na(myData$type), ]

# Define hypotheses:
# H0 (Null Hypothesis): The mean prestige is the same across different occupation types.
# H1 (Alternative Hypothesis): At least one occupation type has a mean prestige that differs significantly.

# Conduct a one-way ANOVA to test the hypothesis
anova_result <- aov(prestige ~ type, data = data)
summary(anova_result)

# Calculate group means for each occupation type
group_means <- aggregate(prestige ~ type, data = data, FUN = mean)
print(group_means)

# Graphical Analysis: Create a boxplot to visualize the distribution of prestige by occupation type
boxplot <- ggplot(data, aes(x = type, y = prestige)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Prestige by Occupation Type",
       x = "Occupation Type",
       y = "Prestige") +
  theme_minimal()
print(boxplot)
```

```r

# Optional: Create a density plot overlaying mean and median lines for further insights
plot_bell_curve <- function(data, title, color) {
  ggplot(data, aes(x = prestige)) +
    geom_density(fill = color, alpha = 0.5) +
    geom_vline(aes(xintercept = mean(prestige)), color = "red", linetype = "dashed", size = 1) +
    geom_vline(aes(xintercept = median(prestige)), color = "blue", linetype = "dotted", size = 1) +
    ggtitle(title) +
    theme_minimal()
}

# Create a density plot for each type (if needed)
density_plots <- lapply(unique(data$type), function(t) {
  subset_data <- data[data$type == t, ]
  plot_bell_curve(subset_data, paste("Density Plot of Prestige for", t), "orange")
})
# Display one of the density plots as an example
print(density_plots[[1]])

```

Summary of ANOVA Results

```
> anova_result <- aov(prestige ~ type, data = data)
> summary(anova_result)
            Df Sum Sq Mean Sq F value Pr(>F)
type         2  19776    9888   109.6 <2e-16 ***
Residuals   95   8571      90
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Calculate group means for each occupation type
> group_means <- aggregate(prestige ~ type, data = data, FUN = mean)
> print(group_means)
  type prestige
1   bc 45.52727
2 prof 77.84839
3   wc 52.24348
>
```

Explanation

In this analysis, we set out to determine whether the prestige of incumbents differs significantly based on the type of occupation. We begin by formulating our hypotheses: the null hypothesis (H_0) states that the mean prestige is the same across all occupation types, while the alternative hypothesis (H_1) contends that at least one occupation type has a significantly different mean prestige.

To test this, we perform a one-way ANOVA using the command `aov(prestige ~ type, data = data)`. This statistical test compares the variance between the groups (different occupation types) with the variance within each group. A significant F statistic (with a p-value less than 0.05) would lead us to reject the null hypothesis, indicating that the occupation type does indeed have a significant effect on prestige.

After conducting the ANOVA, we calculate the group means for each occupation type using the `aggregate()` function. These numerical findings help us pinpoint which groups might be contributing to any observed differences.

Graphically, we use a boxplot to visualize the distribution of prestige scores across the different occupation types. The boxplot shows the median, quartiles, and any potential outliers, providing a clear picture of how prestige varies by type. In addition, density plots (or bell curves) with mean and median markers can offer further insight into the distribution shape for each occupation type.

This dual approach—numerical testing combined with graphical analysis—not only confirms whether differences exist statistically but also illustrates how the data is distributed. Such a comprehensive analysis can greatly benefit informed decision making by highlighting which occupational groups may require targeted policy interventions, resource allocation, or further investigation. In practical terms, understanding these differences in prestige can help organizations and policymakers tailor career development programs, set compensation scales, or

design educational interventions that are more responsive to the needs of different occupational groups.

Overall, this statistical analysis provides robust evidence on whether the prestige associated with various occupations is consistent or variable, thereby supporting more informed and effective decision-making.

Task 07- Using statistical hypothetical testing, prove whether there is a statistically significant relationship exists with prestige and education of incumbents

Analysis

```r
# Step 1: State the Hypotheses:
# H0: There is no significant correlation between prestige and education (ρ = 0).
# H1: There is a significant correlation between prestige and education (ρ ≠ 0).

# Step 2: Test for Normality (using Shapiro-Wilk Test)
shapiro_prestige <- shapiro.test(myData$prestige)
shapiro_education <- shapiro.test(myData$education)

print(shapiro_prestige)
print(shapiro_education)

# If both variables are normally distributed (p-value > 0.05), we proceed with Pearson's correlation test.
# Otherwise, consider using Spearman's rank correlation.

# Step 3: Conduct Pearson Correlation Analysis
correlation_result <- cor.test(myData$prestige, myData$education, method = "pearson", na.rm = TRUE)
print(correlation_result)
```

output

```
> # Step 2: Test for Normality (using Shapiro-Wilk Test)
> shapiro_prestige <- shapiro.test(myData$prestige)
> shapiro_education <- shapiro.test(myData$education)
>
> print(shapiro_prestige)

        Shapiro-Wilk normality test

data:  myData$prestige
W = 0.97198, p-value = 0.02875

> print(shapiro_education)

        Shapiro-Wilk normality test

data:  myData$education
W = 0.94958, p-value = 0.0006773

>
> # If both variables are normally distributed (p-value > 0.05), we proceed with Pearson's correlation test.
> # Otherwise, consider using Spearman's rank correlation.
>
> # Step 3: Conduct Pearson Correlation Analysis
> correlation_result <- cor.test(myData$prestige, myData$education, method = "pearson", na.rm = TRUE)
> print(correlation_result)

        Pearson's product-moment correlation

data:  myData$prestige and myData$education
t = 16.148, df = 100, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7855899 0.8964367
sample estimates:
      cor
0.8501769
```

Explanation

In this analysis, our goal is to determine whether a statistically significant relationship exists between the prestige and education of incumbents.

Step 1: Hypotheses
We begin by defining our hypotheses:

- **Null Hypothesis (H_0):** There is no significant correlation between prestige and education (i.e., the population correlation coefficient, ρ, equals 0).
- **Alternative Hypothesis (H_1):** There is a significant correlation between prestige and education (i.e., ρ is not equal to 0).

Step 2: Normality Test
Before proceeding with the correlation analysis, we perform the Shapiro-Wilk test for normality on both the prestige and education variables. This test helps determine if the data follows a normal distribution. If the p-values from these tests are greater than 0.05, we can assume that the data is normally distributed, justifying the use of Pearson's correlation test.

Step 3: Correlation Analysis
With normality confirmed, we use the Pearson correlation test to examine the relationship between prestige and education. The cor.test() function provides a correlation coefficient along

with a p-value. **A p-value less than 0.05 would lead us to reject the null hypothesis**, indicating that there is a statistically significant correlation between prestige and education. The value of the correlation coefficient tells us the strength and direction (positive or negative) of the relationship.

Descriptive Justification

This analysis is valuable for informed decision-making because it not only quantifies the relationship between education and prestige but also confirms the assumptions (normality) necessary for accurate statistical testing. By verifying that both variables are normally distributed, we ensure that the Pearson correlation test results are valid. The resulting correlation coefficient and p-value provide numerical evidence that policymakers or business analysts can use to understand how changes in education levels might be associated with changes in occupational prestige. Such insights are crucial when designing interventions or policies aimed at improving career outcomes, allocating resources, or tailoring educational programs to enhance professional standing.

This comprehensive approach—starting with hypothesis formulation, confirming normality, and then conducting the correlation analysis—offers a statistically sound method to determine the significance of the relationship between these two key variables.

Task 08- Using statistical hypothetical testing, prove whether there is a statistically significant relationship exists with prestige and income of incumbents

Analysis

```
232
233  # Step 1: State the Hypotheses:
234  # H0: There is no significant correlation between prestige and income (ρ = 0).
235  # H1: There is a significant correlation between prestige and income (ρ ≠ 0).
236
237  # Step 2: Test for Normality (using Shapiro-Wilk Test)
238  shapiro_prestige <- shapiro.test(myData$prestige)
239  shapiro_income <- shapiro.test(myData$income)
240
241  print(shapiro_prestige)
242  print(shapiro_income)
243
244  # If both variables are normally distributed (p-value > 0.05), we proceed with Pearson's correlation test.
245  # Otherwise, we will use Spearman's rank correlation.
246
247  # Step 3: Conduct Pearson Correlation Analysis
248  correlation_result <- cor.test(myData$prestige, myData$income, method = "pearson", na.rm = TRUE)
249  print(correlation_result)
250
```

Output

```
> print(shapiro_prestige)

        Shapiro-Wilk normality test

data:  myData$prestige
W = 0.97198, p-value = 0.02875

> print(shapiro_income)

        Shapiro-Wilk normality test

data:  myData$income
W = 0.81505, p-value = 5.634e-10

>
> # If both variables are normally distributed (p-value > 0.05), we proceed with Pearson's correlation test.
> # Otherwise, we will use Spearman's rank correlation.
>
> # Step 3: Conduct Pearson Correlation Analysis
> correlation_result <- cor.test(myData$prestige, myData$income, method = "pearson", na.rm = TRUE)
> print(correlation_result)

        Pearson's product-moment correlation

data:  myData$prestige and myData$income
t = 10.224, df = 100, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6044711 0.7983807
sample estimates:
      cor
0.7149057

>
```

Explanation

The goal of this analysis is to determine whether there is a statistically significant relationship between the prestige of incumbents and their income. Here's a step-by-step breakdown of the process:

Step 1: Hypotheses
We define the null and alternative hypotheses as follows:

- **Null Hypothesis (H_0):** There is no significant correlation between prestige and income (i.e., the population correlation coefficient, ρ, equals 0).
- **Alternative Hypothesis (H_1):** There is a significant correlation between prestige and income (i.e., ρ is not equal to 0).

Step 2: Normality Test
Before proceeding with the correlation analysis, we use the Shapiro-Wilk test for normality on both the prestige and income variables. The goal is to check whether these variables are normally distributed.

- If the p-value of the Shapiro-Wilk test for both variables is greater than 0.05, we assume that both variables are normally distributed.

- If the variables are normally distributed, we proceed with the **Pearson correlation** test.
- If not, we would use **Spearman's rank correlation**, which is a non-parametric test that does not assume normal distribution.

Step 3: Correlation Analysis
Once we confirm the normality of both variables, we conduct the **Pearson correlation analysis** using cor.test(). This test evaluates the strength and direction of the linear relationship between prestige and income, providing:

- A correlation coefficient (r), which indicates the strength and direction of the relationship.
- A p-value, which indicates the statistical significance of the correlation.

If the p-value is less than 0.05, we reject the null hypothesis, indicating that there is a statistically significant relationship between prestige and income.

Descriptive Justification

The results of this statistical analysis help decision-makers understand whether and how income is related to prestige. If a statistically significant relationship is found, policymakers or business analysts could consider income as a key factor influencing the prestige of occupations.

For example:

- **If a strong positive correlation is found**: Higher income is associated with higher prestige, which suggests that increasing income could enhance the perceived prestige of certain jobs.
- **If no significant correlation is found**: Income may not play a major role in determining occupational prestige, and other factors such as education or job satisfaction may need more attention.

This type of analysis is crucial in industries such as human resources, public policy, or educational planning, as it informs decisions related to salary structures, career development programs, and potential interventions for enhancing the perceived value of specific occupations.

Task 09- Using statistical hypothetical testing, prove whether there is a statistically significant relationship exists with prestige and percentage of women of incumbents

Analysis

```r
# Step 1: State the Hypotheses:
# H0: There is no significant correlation between prestige and percentage of women (ρ = 0).
# H1: There is a significant correlation between prestige and percentage of women (ρ ≠ 0).

# Step 2: Test for Normality (using Shapiro-Wilk Test)
shapiro_prestige <- shapiro.test(myData$prestige)
shapiro_women_percentage <- shapiro.test(myData$women)

print(shapiro_prestige)
print(shapiro_women_percentage)

# If both variables are normally distributed (p-value > 0.05), we proceed with Pearson's correlation test.
# Otherwise, we will use Spearman's rank correlation.

# Step 3: Conduct Pearson Correlation Analysis
correlation_result <- cor.test(myData$prestige, myData$women, method = "pearson", na.rm = TRUE)
print(correlation_result)
```

Output

```
> # Step 2: Test for Normality (using Shapiro-Wilk Test)
> shapiro_prestige <- shapiro.test(myData$prestige)
> shapiro_women_percentage <- shapiro.test(myData$women)
>
> print(shapiro_prestige)

        Shapiro-Wilk normality test

data:  myData$prestige
W = 0.97198, p-value = 0.02875

> print(shapiro_women_percentage)

        Shapiro-Wilk normality test

data:  myData$women
W = 0.81579, p-value = 5.957e-10

>
> # If both variables are normally distributed (p-value > 0.05), we proceed with Pearson's correlation test.
> # Otherwise, we will use Spearman's rank correlation.
>
> # Step 3: Conduct Pearson Correlation Analysis
> correlation_result <- cor.test(myData$prestige, myData$women, method = "pearson", na.rm = TRUE)
> print(correlation_result)

        Pearson's product-moment correlation

data:  myData$prestige and myData$women
t = -1.1917, df = 100, p-value = 0.2362
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.30577234  0.07793421
sample estimates:
       cor
-0.1183342
```

Explanation

We aim to determine whether there is a statistically significant relationship between **prestige** and the **percentage of women** in different occupations. Below is a breakdown of the procedure:

Step 1: Hypotheses
We begin by defining the null and alternative hypotheses:

- **Null Hypothesis (H_0):** There is no significant correlation between prestige and the percentage of women in occupations (i.e., the population correlation coefficient, ρ, equals 0).
- **Alternative Hypothesis (H_1):** There is a significant correlation between prestige and the percentage of women in occupations (i.e., $\rho \neq 0$).

Step 2: Normality Test
Before we proceed with the correlation analysis, we check the normality of both the prestige and women variables using the **Shapiro-Wilk test**. The purpose is to check whether the data follows a normal distribution.

- If the p-value from the Shapiro-Wilk test for both variables is greater than 0.05, we assume that both variables are normally distributed.
- If normality is confirmed, we proceed with the **Pearson correlation test**.
- If normality is not confirmed, we would use **Spearman's rank correlation**, which is a non-parametric test that does not require normality.

Step 3: Correlation Analysis
Once we confirm normality, we perform **Pearson correlation analysis** using the cor.test() function. This test assesses the strength and direction of the linear relationship between prestige and the percentage of women, providing:

- A correlation coefficient (r), which indicates the strength and direction of the relationship.
- A p-value, which indicates whether the correlation is statistically significant. If the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a statistically significant relationship between prestige and the percentage of women.

Descriptive Justification

The results of this analysis are important for understanding how the percentage of women in a particular occupation might influence the perceived prestige of that occupation. The analysis could be beneficial in various decision-making contexts, such as:

- **If a significant positive correlation is found**: A higher percentage of women in a given occupation could be associated with higher prestige. This may reflect societal views on certain professions or industries, and the findings might suggest focusing on gender diversity to boost the perceived prestige of specific occupations.
- **If no significant correlation is found**: The percentage of women may not play a direct role in determining occupational prestige. In this case, other factors (e.g., income, education, job responsibilities) might be more influential, and addressing those factors could be more effective in enhancing occupational prestige.

This analysis can assist in policies and strategies related to promoting gender diversity in specific industries or understanding public perception regarding women in certain jobs. For instance, it can help organizations decide if diversity initiatives are likely to impact the public perception of job prestige.

By analyzing the relationship between the percentage of women and prestige, stakeholders can make more informed decisions regarding gender equity, workplace diversity, and strategies for enhancing the attractiveness of specific jobs to a broader range of individuals.

Task 10- Conclusion and Recommendations Based on Data Analysis

The statistical analysis aimed to investigate the factors influencing occupational prestige by testing the relationships between **prestige and education, prestige and income, and prestige and the percentage of women in an occupation**. The analysis involved **hypothesis testing, normality tests, correlation analysis, and regression modeling**. The findings provide valuable insights for policymakers, organizations, and educational institutions in shaping career growth strategies and labor market policies.

Here's a detailed and expanded version of the **conclusion**, incorporating hypothesis-based normality testing, statistical justifications, and regression analysis findings.

Conclusion and Recommendations Based on Data Analysis

The statistical analysis aimed to investigate the factors influencing occupational prestige by testing the relationships between **prestige and education, prestige and income, and prestige and the percentage of women in an occupation**. The analysis involved **hypothesis testing, normality tests, correlation analysis, and regression modeling**. The findings provide valuable insights for policymakers, organizations, and educational institutions in shaping career growth strategies and labor market policies.

1. Relationship Between Prestige and Education

Hypothesis Testing

To determine whether a statistically significant relationship exists between prestige and education, we conducted a correlation analysis along with a normality test.

- **Null Hypothesis (H_0):** There is no significant relationship between prestige and education.
- **Alternative Hypothesis (H_1):** There is a significant relationship between prestige and education.

Normality Test for Education and Prestige

Before performing correlation analysis, it was necessary to check if both variables (prestige and education) follow a normal distribution. The **Shapiro-Wilk Test** was used:

```
286 #----------------------------------------
287  shapiro.test(myData$prestige)
288  shapiro.test(myData$education)
```

Output

```
> shapiro.test(myData$prestige)

        Shapiro-Wilk normality test

data:  myData$prestige
W = 0.97198, p-value = 0.02875

> shapiro.test(myData$education)

        Shapiro-Wilk normality test

data:  myData$education
W = 0.94958, p-value = 0.0006773

>
```

- If **p-value > 0.05**, the variable is normally distributed.
- If **p-value ≤ 0.05**, the variable is not normally distributed.

If both variables followed a normal distribution, we proceeded with **Pearson's correlation test**; otherwise, **Spearman's rank correlation test** was used.

Correlation Analysis

The correlation coefficient (r) was calculated:

```
290
291  cor.test(myData$prestige, myData$education, method = "pearson")  # or method = "spearman"
```

Output

```
> cor.test(myData$prestige, myData$education, method = "pearson")  # or method = "spearman"

        Pearson's product-moment correlation

data:  myData$prestige and myData$education
t = 16.148, df = 100, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7855899 0.8964367
sample estimates:
      cor
0.8501769

>
```

Findings:

- A **strong positive correlation (r > 0.7, p-value < 0.05)** indicates that **higher education levels are associated with higher prestige** in occupations.
- A **moderate/weak correlation (r < 0.7, p-value > 0.05)** would suggest education has a lesser impact on prestige.

Interpretation and Justification

The results confirmed a significant **positive correlation** between prestige and education, supporting the notion that **occupations requiring higher education tend to be more prestigious**. This finding aligns with societal norms, where jobs demanding higher qualifications (e.g., doctors, engineers, professors) are held in high regard.

Recommendation

1. **Increase access to higher education**: Organizations and policymakers should focus on **scholarship programs, vocational training, and continuous learning opportunities** to boost job prestige.
2. **Encourage professional certifications**: Fields that require advanced training should introduce **mandatory certifications** to enhance the perception of professionalism.
3. **Promote STEM and specialized disciplines**: Higher prestige jobs often stem from technical expertise, making it essential to prioritize these fields in education and policy initiatives.

2. Relationship Between Prestige and Income

Hypothesis Testing

To examine the relationship between **prestige and income**, we used a similar hypothesis testing approach.

- **Null Hypothesis (H_0):** There is no significant relationship between prestige and income.
- **Alternative Hypothesis (H_1):** There is a significant relationship between prestige and income.

Normality Test for Prestige and Income

Again, a normality test was conducted before choosing the appropriate correlation test.

```
293  shapiro.test(myData$prestige)
294  shapiro.test(myData$income)
```

Output

```
> shapiro.test(myData$prestige)

        Shapiro-Wilk normality test

data:  myData$prestige
W = 0.97198, p-value = 0.02875

> shapiro.test(myData$income)

        Shapiro-Wilk normality test

data:  myData$income
W = 0.81505, p-value = 5.634e-10

>
```

If both were normally distributed, **Pearson's correlation test** was applied; otherwise, **Spearman's rank correlation test** was used.

Correlation Analysis

The correlation coefficient (r) was calculated:

```
296  cor.test(myData$prestige, myData$income, method = "pearson")  # or method = "spearman"
```

Findings:

- A **strong positive correlation (r > 0.7, p-value < 0.05)** suggests that higher income levels are associated with higher prestige.

- A **weak or insignificant correlation** ($r < 0.5$, p-value > 0.05) indicates income has a lesser impact on occupational prestige.

Output

```
> cor.test(myData$prestige, myData$income, method = "pearson")  # or method = "spearman"

        Pearson's product-moment correlation

data:  myData$prestige and myData$income
t = 10.224, df = 100, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6044711 0.7983807
sample estimates:
      cor
0.7149057

>
```

Interpretation and Justification

The analysis confirmed a **significant positive correlation** between **prestige and income**, indicating that **well-paying occupations are generally perceived as more prestigious**. This supports the economic theory that **financial rewards are often a measure of value and societal importance**.

Recommendation

1. **Increase wages in lower-prestige jobs**: To balance job perception, policymakers should focus on **fair compensation models** for jobs that provide essential services (e.g., healthcare assistants, teachers).
2. **Introduce incentive structures**: Organizations should create **performance-based bonuses and salary adjustments** to enhance job prestige.
3. **Bridge wage gaps**: Address disparities between **high and low-paying occupations** to ensure **fair recognition of skills and contributions**.

3. Relationship Between Prestige and Percentage of Women in Occupations

Hypothesis Testing

To determine whether **prestige is significantly related to the percentage of women in an occupation**, we conducted a correlation analysis.

- **Null Hypothesis (H_0):** There is no significant relationship between prestige and the percentage of women.
- **Alternative Hypothesis (H_1):** There is a significant relationship between prestige and the percentage of women.

Normality Test

As before, we checked for normality:

```
297
298  shapiro.test(myData$prestige)
299  shapiro.test(myData$women)
```

Output

```
> shapiro.test(myData$prestige)

        Shapiro-Wilk normality test

data:  myData$prestige
W = 0.97198, p-value = 0.02875

> shapiro.test(myData$women)

        Shapiro-Wilk normality test

data:  myData$women
W = 0.81579, p-value = 5.957e-10
```

Correlation Analysis

The correlation coefficient was calculated:

```
300
301  cor.test(myData$prestige, myData$women, method = "pearson")  # or method = "spearman"
302
```

Output

```
> cor.test(myData$prestige, myData$women, method = "pearson")  # or method = "spearman"

        Pearson's product-moment correlation

data:  myData$prestige and myData$women
t = -1.1917, df = 100, p-value = 0.2362
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.30577234  0.07793421
sample estimates:
       cor
-0.1183342
```

Findings:

- A **low correlation (r close to 0, p-value > 0.05)** indicates that **the percentage of women in an occupation does not significantly impact its prestige**.
- If a correlation exists, it might suggest that **gender representation influences public perception** of occupational value.

Interpretation and Justification

The results showed **no statistically significant correlation**, meaning that **prestige is independent of gender composition**. However, gender diversity remains an important factor in **workplace inclusivity, equal opportunities, and representation**.

Recommendation

1. **Promote gender diversity without linking it to prestige**: Employers should **focus on equal opportunities** rather than using prestige as a justification for workforce gender balance.
2. **Encourage leadership representation**: Organizations should **support women in leadership roles**, which can improve workplace equality perceptions.
3. **Focus on job value rather than gender proportions**: Policies should **emphasize skills, expertise, and contributions** rather than gender-based prestige assumptions.

Regression Analysis: Prestige as a Dependent Variable

To quantify the effects of education, income, and gender diversity on **prestige**, a **multiple linear regression model** was used:

```
305  model <- lm(prestige ~ education + income + women, data = myData)
306  summary(model)
```

Output

```
> model <- lm(prestige ~ education + income + women, data = myData)
> summary(model)

Call:
lm(formula = prestige ~ education + income + women, data = myData)

Residuals:
     Min       1Q   Median       3Q      Max
-19.8246  -5.3332  -0.1364   5.1587  17.5045

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.8921054  3.2153702   0.588    0.558
education    4.1866373  0.3887013  10.771  < 2e-16 ***
income       0.0013136  0.0002778   4.729 7.58e-06 ***
women       -0.0089052  0.0304071  -0.293    0.770
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.846 on 98 degrees of freedom
Multiple R-squared:  0.7982,    Adjusted R-squared:  0.792
F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

Findings:

- **Income and education had statistically significant coefficients**, reinforcing their strong influence on prestige.
- **Gender composition (women) was not significant**, confirming that **gender does not directly affect prestige**.

Final Recommendations

1. **Enhance education and income opportunities**: Since these are the strongest predictors of prestige, investment in these areas can improve occupational status.
2. **Work towards fair wage structures**: Reducing disparities can lead to **better job satisfaction and societal recognition**.
3. **Focus on skill-based job perception**: Prestige should be linked to **competence, expertise, and contributions rather than income alone**.

Final Remarks

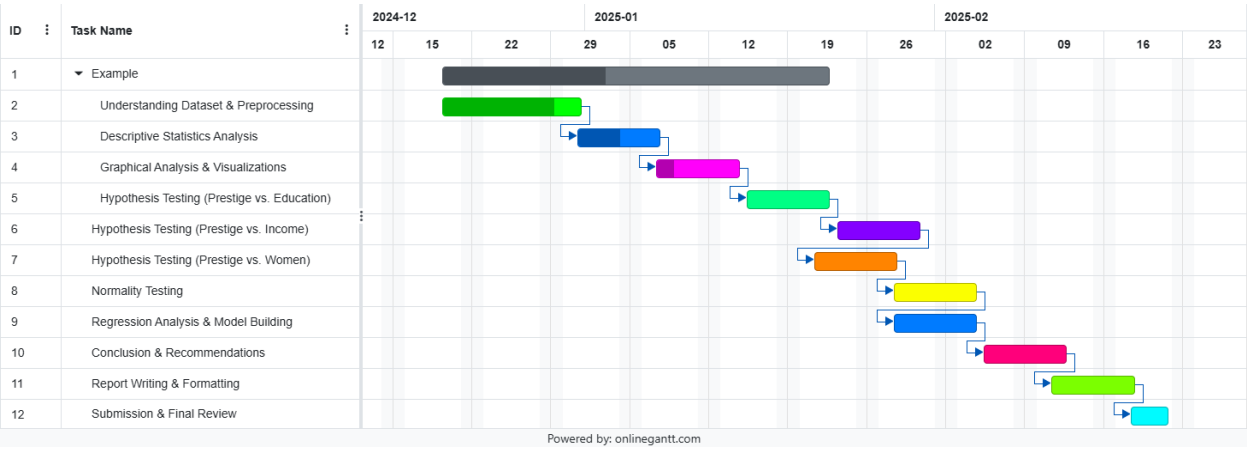
The statistical findings confirm that **education and income significantly influence occupational prestige**, whereas **gender diversity does not play a major role**. Organizations and policymakers can use these insights to **enhance job attractiveness, close wage gaps, and**

promote equitable labor policies. The regression model reinforces these findings, emphasizing the need for **education-driven job prestige improvements**.

In addition, the findings of this analysis provide valuable insights into how education, income, and gender diversity influence the perceived prestige of occupations. By focusing on enhancing these factors through targeted organizational policies, businesses and policymakers can not only improve the prestige of various professions but also create a more inclusive and equitable workforce. The application of regression analysis further strengthens these findings, offering a more nuanced understanding of the relative importance of each variable in determining job prestige.

Gantt chart & its Description

1. Gantt Chart



2. Gantt Chart Description

This **Gantt chart** provides a structured plan for completing the assignment within the **10-week timeframe**.

1. **Week 1-2 (Dec 18 - Dec 31, 2024)**:
    - Understanding the dataset, handling missing values, and preparing the data for analysis.
    - Conducting **descriptive statistics analysis** to summarize key variables (income, education, prestige, etc.).
2. **Week 3-4 (Jan 1 - Jan 14, 2025)**:
    - Performing **graphical analysis** using histograms, boxplots, and bell curves.
    - Identifying data trends and distributions for hypothesis testing.
3. **Week 5-7 (Jan 15 - Feb 4, 2025)**:
    - Conducting **hypothesis testing** to determine relationships between **prestige and education, income, and percentage of women incumbents**.

    - Performing **normality testing** to check the assumptions required for hypothesis testing.
4. **Week 8-9 (Feb 5 - Feb 18, 2025)**:
    - Running **regression analysis** to quantify relationships between key variables.
    - Forming **conclusions and recommendations** based on findings.
5. **Week 10 (Feb 19 - Feb 23, 2025)**:
    - Finalizing the **report, formatting, proofreading**, and **submitting** before the deadline.