**Title:** Enterprise Business Analytics of Young Teenage Affairs in Eastern region of Uganda.

1. Introduction

**Objective of the assignment**

To collect, clean, integrate, analyze, visualize and document young teenage affairs dataset for better decision making in businesses.

**Description of the dataset**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **gender** | factor indicating gender. |  |  |  |  |  |  |
| **age** | numeric variable coding age in years: 17.5 = under 20, 22 = 20–24, 27 = 25–29,  32 = 30–34, 37 = 35–39,42 = 40–44, 47 = 45–49, 52 = 50–54, 57 = 55 or over. | | | | |  |  |
| **yearsmarried** | numeric variable coding number of years married: 0.125 = 3 months or less,  0.417 = 4–6 months, 0.75 = 6 months–1 year,1.5 = 1–2 years, 4 = 3–5 years,  7 = 6–8 years, 10 = 9–11 years, 15 = 12 or more years. | | | | | | |
| **children** | factor. Are there children in the marriage? |  |  |  |  |  |  |
| **religiousness** | numeric variable coding religiousness: 1 = anti, 2 = not at all, 3 = slightly,  4 = somewhat, 5 = very. |  |  |  |  |  |  |
| **education** | numeric variable coding level of education: 9 = grade school, 12 = high school graduate,  14 = some college, 16 = college graduate, 17 = some graduate work,  18 = master's degree, 20 = Ph.D., M.D., or other advanced degree. | | | | | | |
| **occupation** | numeric variable coding occupation according to Hollingshead classification (reverse numbering). |  |  |  |  |  |  |
| **rating** | numeric variable coding self rating of marriage: 1 = very unhappy, 2 = somewhat unhappy,  3 = average, 4 = happier than average, 5 = very happy. | | |  |  |  |  |

1. **Data Collection**

**Data sources**

The web/ internet.

**Data collection methods**

The data set is a downloaded dataset of young teenage affairs.

1. **Data Cleaning**

**Data cleaning steps**

> library(readxl)

> Big\_data

<- read\_excel("C:/Users/atwii/OneDrive/Desktop/EBA-TH 2024/Big data.xlsx")  
> View(Big\_data)

> library(readxl)

> Small\_data

<- read\_excel("C:/Users/atwii/OneDrive/Desktop/EBA-TH 2024/Small data.xlsx")> View(Small\_data)

> head(Big\_data)

# View the first few rows of the dataset

head(Big\_data)

head(Small\_data)

# Summary of the dataset

summary(Big\_data)

# Check the structure of the dataset

str(Big\_data)

# Displaying the names of the columns

names(Big\_data)

# Checking for missing values

sum(is.na(Big\_data))

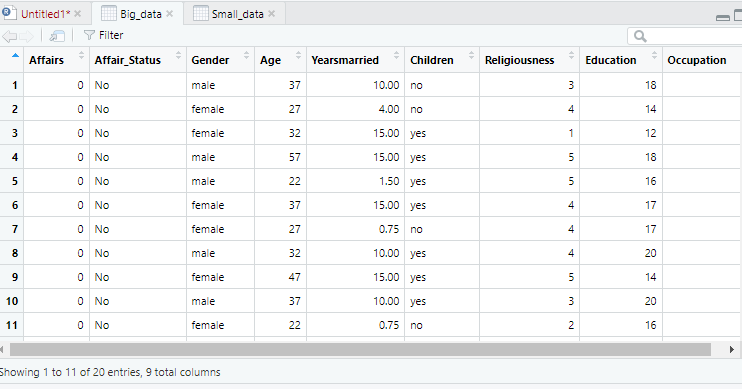
# Removing rows with missing values

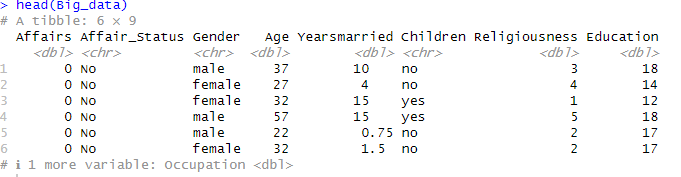
Big\_data\_cleaned <- na.omit(“column1”,”Rating”)

# Impute missing values (e.g., using the mean for numeric columns)

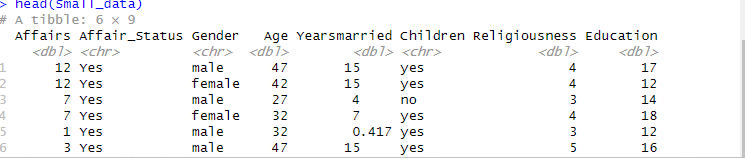
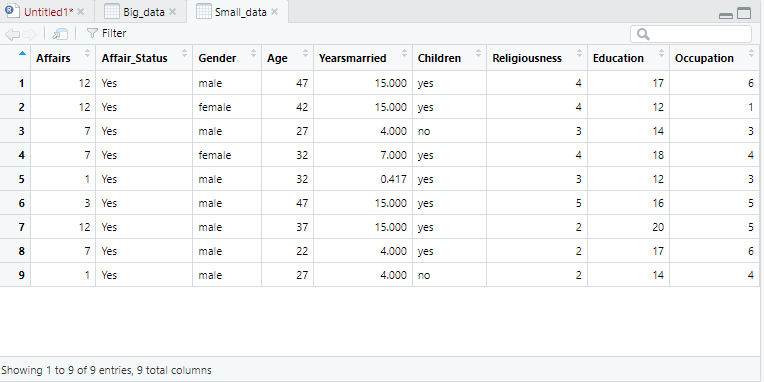
Big\_data$Rating[is.na(Big\_data$Rating)] <- mean(Big\_data$Rating, na.rm = TRUE)

**Before cleaning**





**After cleaning**



1. **Data Integration**

**Additional data sources**

Small data sets of young teenagers with affairs and having affairs status as “yes”

**Data integration process**

# Load datasets

Big\_data <- read\_excel("Big\_data.xlsx")

Small\_data <- read\_excel("Small\_data.xlsx")

combined\_data <- intergrated %>%

left\_join(Big\_data, by = "Affairs",) %>%

left\_join(Small\_data, by = "Affairs",)

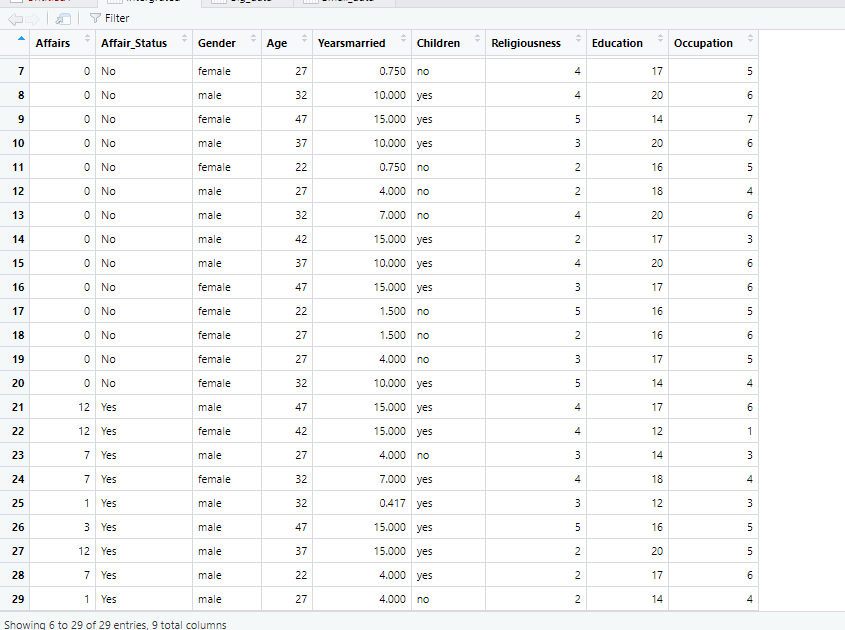
> combined\_data <- rbind(Big\_data, Small\_data)

> library(readxl)> intergrated

<- read\_excel("C:/Users/atwii/OneDrive/Desktop/EBA-TH 2024/intergrated.xlsx")

> View(intergrated)

**Snapshot:**



**Reasons for Integrating data sources**

**1. Comprehensive Understanding**

**Objective:** To obtain a holistic view of young teenage affairs by combining data from various sources.

**Rationale**

Young teenage affairs are multifaceted, and understanding them requires insights from different aspects such as educational performance, health statistics, social interactions, and family background. Integrating diverse datasets helps in creating a comprehensive picture of the issues affecting teenagers.

**2. Enhanced Data Quality and Accuracy**

**Objective:** To improve the reliability and validity of the analysis.

**Rationale**

Data from a single source may have limitations or biases. Integrating multiple sources can help cross-verify information, fill in gaps, and correct errors. This results in more accurate and robust findings.

**3. Identifying Patterns and Trends**

**Objective:**To discover patterns and trends that are not visible in isolated datasets.

**Rationale:**

Different datasets can reveal trends and correlations that are not apparent when viewed separately. Integration allows for the examination of relationships between variables across different domains.

**4. Policy and Intervention Design**

**Objective:**To inform policy decisions and design effective interventions.

**Rationale:**

Integrated data provides a more complete view of the issues, which is essential for designing targeted policies and interventions. Policymakers can use integrated data to identify key areas for intervention and allocate resources more effectively.

**5. Data-Driven Insights and Decision-Making**

**Objective:**To base decisions on evidence and data-driven insights rather than anecdotal information.

**Rationale:**

Integrating data from multiple sources ensures that decisions are based on comprehensive and objective evidence. This is crucial for addressing complex issues where multiple factors are at play.

**6. Improving Predictive Models**

**Objective:**To enhance the accuracy of predictive models used in understanding teenage affairs

**Rationale:**

Predictive modeling benefits from rich and diverse datasets. Integrating data from different sources allows for the development of more accurate models that can forecast trends and behaviors.

7**. Efficiency and Resource Optimization**

**Objective:**To streamline data analysis and optimize resource use.

**Rationale:**

Integrating data sources can reduce the need for redundant data collection and analysis, leading to more efficient use of resources. This is particularly important in research and policy-making contexts where time and budget constraints exist.

**Processes of Data Integration in R**

**1. Loading Data**

The first step is to load data from various sources into R. This can be done using functions from base R or libraries like readr, readxl, or data.table.

**#Load libraries**

library(readr)

library(readxl)

library(data.table)

**# Load datasets**

data1 <- read\_csv("path/to/data1.csv")

data2 <- read\_excel("path/to/data2.xlsx")

**2. Inspecting Data**

Examine the structure and content of each dataset to understand the format, column names, and data types.

# **View the first few rows**

head(data1)

head(data2)

**# Check the structure**

str(data1)

str(data2)

**# Summary statistics**

summary(data1)

summary(data2)

**3. Cleaning and Preparing Data**

Ensure that the datasets are clean and in a consistent format before merging. This includes handling missing values, correcting data types, and standardizing column names.

**# Rename columns for consistency**

names(data1) <- c("ID", "Value1")

names(data2) <- c("ID", "Value2")

**4. Merging Data**

Combine datasets using functions like merge() or join operations from the dplyr package. The choice of merge operation (inner join, left join, right join, full join) depends on the requirements of your analysis.

**# Base R merge**

combined\_data <- merge(data1, data2, by = "ID", all.x = TRUE)

combined\_data <- merge(combined\_data, data3, by = "ID", all.x = TRUE)

**# Using dplyr for joins**

library(dplyr)

combined\_data <- data1 %>%

left\_join(data2, by = "ID") %>%

left\_join(data3, by = "ID")

**5. Handling Duplicate and Conflicting Data**

After merging, check for duplicate rows or conflicting data values. Address these issues by removing duplicates or resolving conflicts.

**# Remove duplicate rows**

combined\_data <- distinct(combined\_data)

**6. Transforming Data**

Perform any necessary transformations, such as aggregating data, creating new variables, or normalizing values.

**# Create a new variable**

combined\_data <- combined\_data %>%

mutate(TotalValue = Value1 + Value2 + Value3)

**7. Validating Integrated Data**

Verify that the integrated dataset is accurate and complete by performing sanity checks and validating with known values or benchmarks.

**# Summary statistics and validation checks**

summary(combined\_data)

**8. Saving the Integrated Data**

Save the integrated dataset for further analysis or reporting.

**# Save as CSV**

write\_csv(combined\_data, "path/to/integrated\_data.csv")

**# Save as RData**

save(combined\_data, file = "path/to/integrated\_data.RData")

**Common Challenges and Solutions**

1. **Inconsistent Data Formats**

**Challenge:**

Data from different sources might have inconsistent formats, making integration difficult.

**Solution:**

Standardize formats before merging, including converting data types, renaming columns, and ensuring consistent units.

**# Standardize formats**

data1$Date <- as.Date(data1$Date, format = "%Y-%m-%d")

data2$Date <- as.Date(data2$Date, format = "%d/%m/%Y")

1. **Missing Data**

**Challenge:**

Datasets may have missing values, which can affect the quality of the integrated data.

**Solution:**

Handle missing values by imputation or removal, depending on the extent and importance of the missing data.

**# Impute missing values**

library(mice)

imputed\_data <- mice(combined\_data, method = 'mean', m = 1)

combined\_data <- complete(imputed\_data)

1. **Data Conflicts**

**Challenge:**

Conflicting information or discrepancies between datasets can arise.

**Solution:**

Investigate the sources of conflicts and decide on a resolution strategy, such as prioritizing certain datasets or performing data reconciliation.

**# Resolving conflicts**

combined\_data <- combined\_data %>%

mutate(ResolvedValue = ifelse(is.na(Value1), Value2, Value1))

1. **Scalability**

**Challenge:**

Integrating very large datasets can be computationally intensive and time-consuming.

**Solution:**

Optimize performance by using efficient data manipulation packages like data.table or working with a subset of data if feasible.

**# Using data.table for large datasets**

library(data.table)

data1 <- fread("large\_data1.csv")

data2 <- fread("large\_data2.csv")

combined\_data <- merge(data1, data2, by = "ID")

1. **Data Privacy and Security**

**Challenge:**

Handling sensitive or personal data requires ensuring privacy and compliance with regulations.

**Solution:**

Apply data anonymization techniques, secure data storage, and ensure compliance with relevant data protection laws (e.g., GDPR).

**# Example of anonymization**

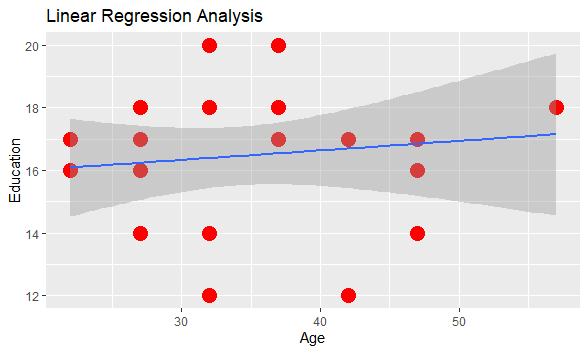
combined\_data$ID <- paste0("ID\_", seq\_along(combined\_data$ID))

1. **Data Analysis**

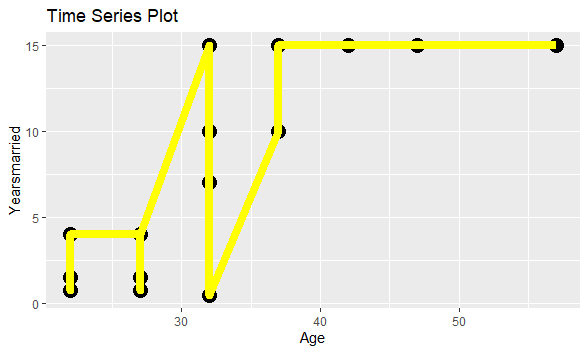
**Exploratory data analysis (EDA)**

Exploratory Data Analysis (EDA) is a crucial step in data analysis that involves summarizing and visualizing datasets to understand their main characteristics, identify patterns, and detect anomalies.

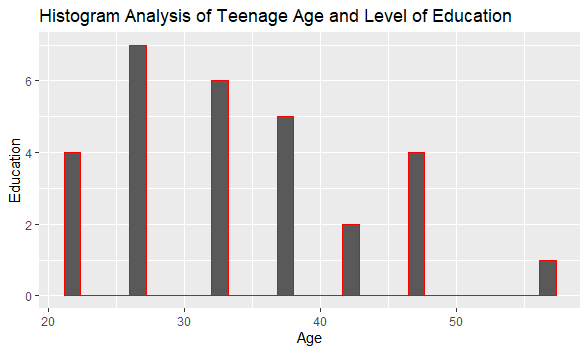
1. **Distribution of variables**

 There is Scatter distribution in the of Teenage level of Education, at the age 30 years they tend to attain more levels but not for long, as they clock 40, they then really drops in Acquiring more education.

1. **Correlations between variables**

 As Teenagers, then to mature from above 30 years, they settle in their affairs hence spending more time together. Rather than when they are below 30 years old.

1. **Patterns or trends observed**

 At The early age between 20 years old to 30 years old, Teenagers tend to acquire more education studies, while as they grow older with more responsibilities of affairs, children, studying does not become a priority.

**Statistical or machine learning analysis**

**Regression Analysis**:

Call:

lm(formula = Education ~ Age, data = intergrated)

Residuals:

Min 1Q Median 3Q Max

-4.6939 -2.2392 0.1545 1.4577 3.6092

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.42080 1.78827 8.623 3.08e-09 \*\*\*

Age 0.03031 0.05101 0.594 0.557

---

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

Residual standard error: 2.457 on 27 degrees of freedom

Multiple R-squared: 0.01291, Adjusted R-squared: -0.02365

F-statistic: 0.3531 on 1 and 27 DF, p-value: 0.5573

**Findings and insights**

C:\Users\user\Desktop\EBA-TH 2024\linear regression.tiff

**Syntax in R**

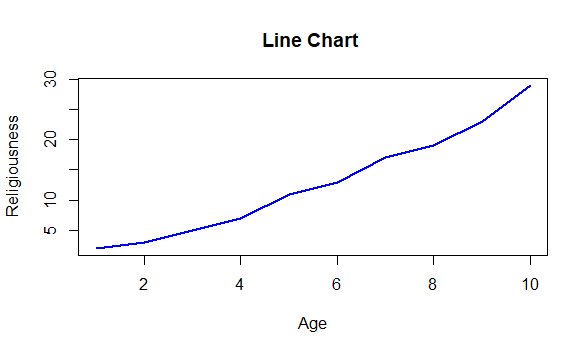
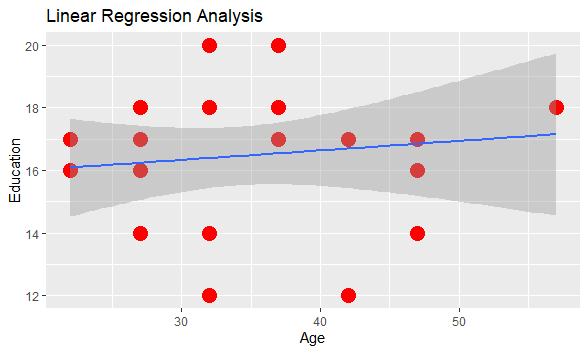
ggplot(intergrated, aes(x = Age, y = Education)) +

geom\_point(size=5,

color="red") +

geom\_smooth(method = "lm") +

labs(title = "Linear Regression Analysis", x = "Age", y = "Education")



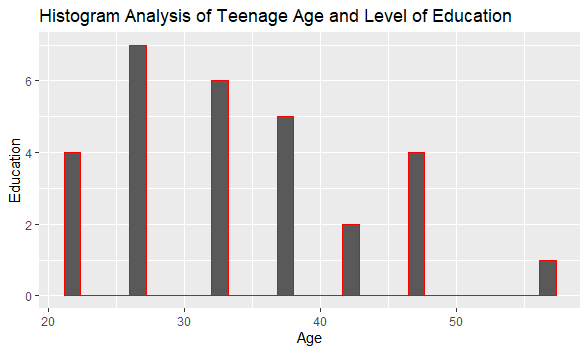
**syntax in R**

ggplot(intergrated, aes(x = Age, )) +

geom\_histogram(bins = 30,

color="red") +

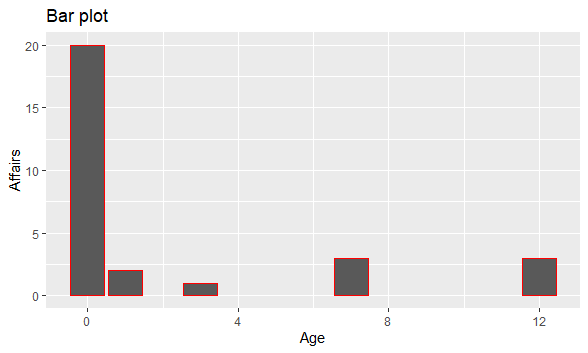
labs(title = "Histogram Analysis of Teenage Age and Level of Education", x = "Age", y = "Education")

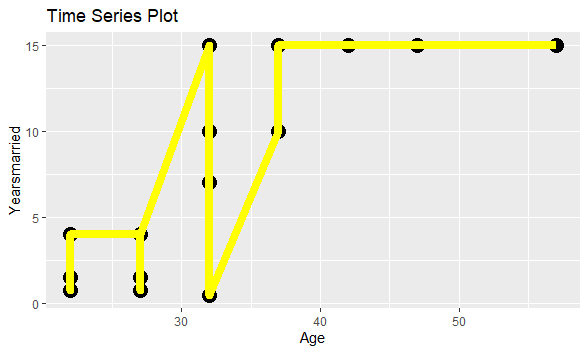


**Bar plots for categorical variables**

ggplot(intergrated, aes(x = Affairs, )) +

geom\_bar(color="red")

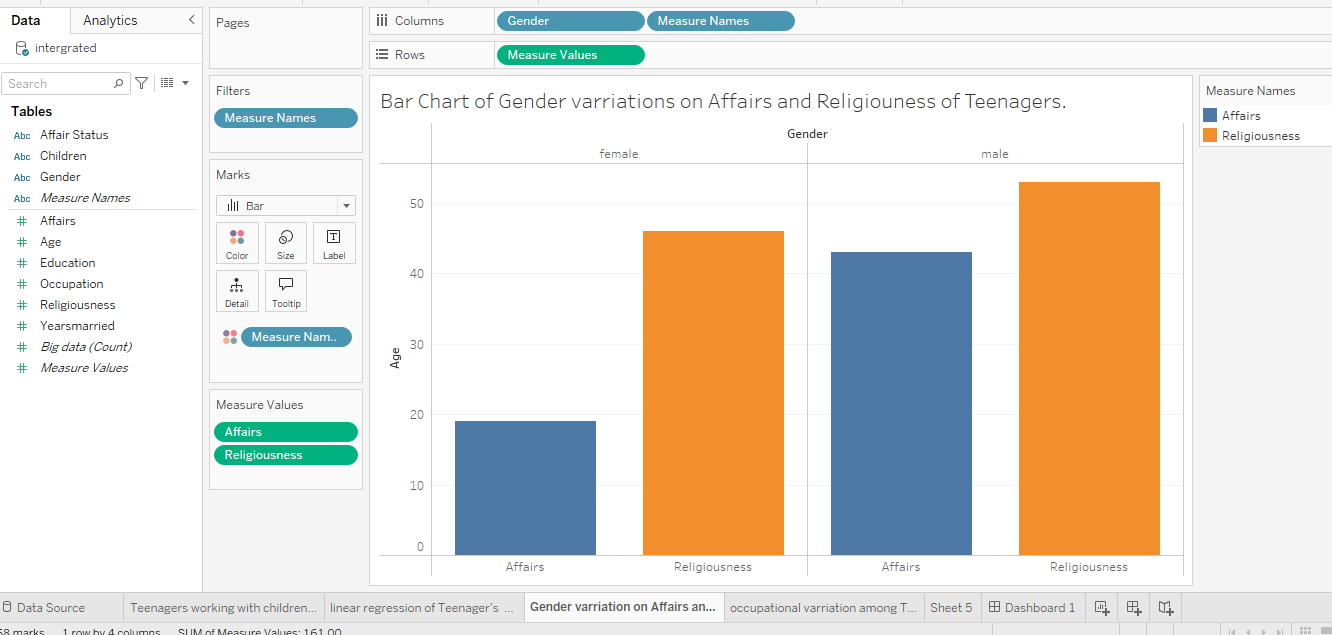
labs(title = "Bar plot", x = "Affairs", y = "Age “)



1. **Data Visualization**

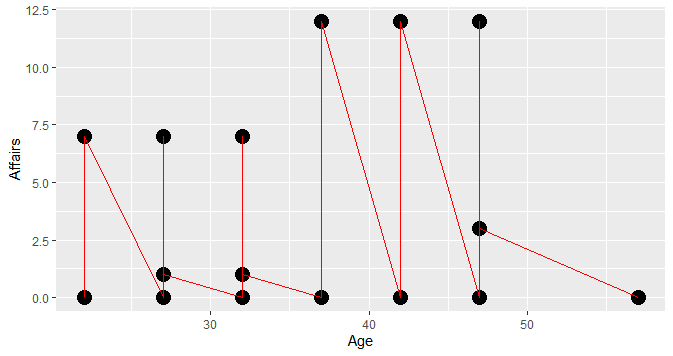
**Description of visualizations**

### ****Bar Charts****



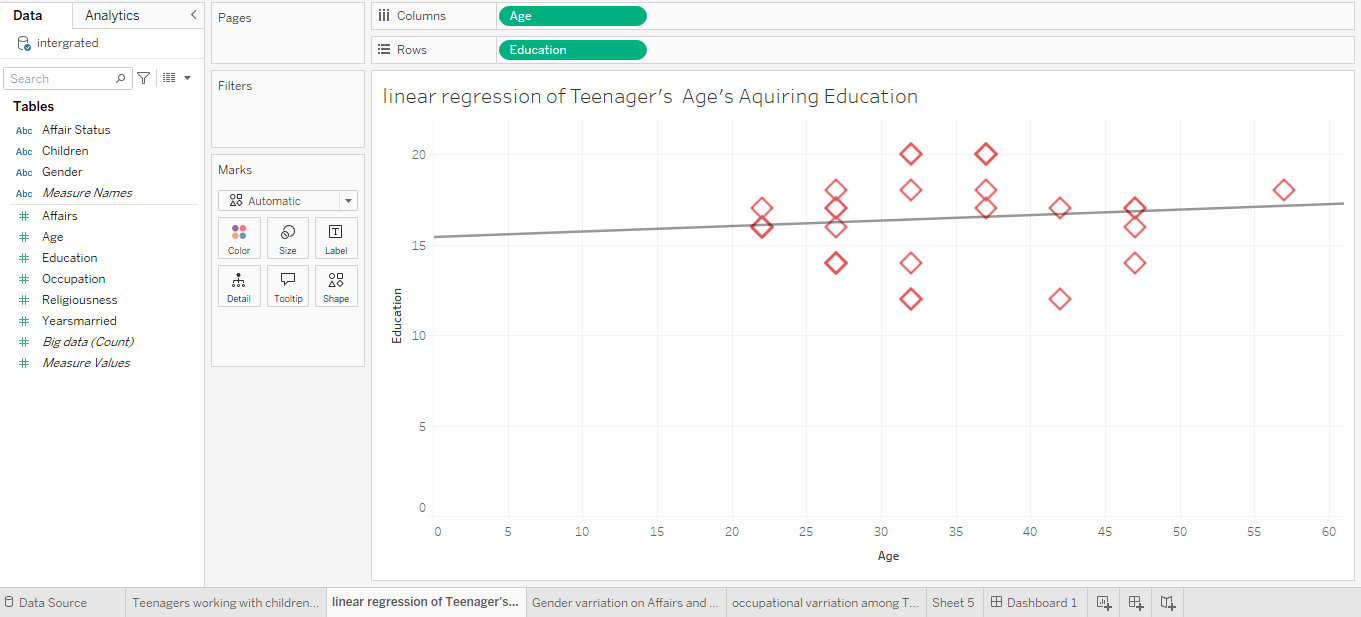
* **Description:** This displays data using rectangular bars with lengths proportional to the values they represent. Each bar represents a category or group, such as Teenager having affairs and them being religious at the different stages of their life.
* **Use:** To compare the frequency or prevalence of teenage affairs across different categories of Affairs and Religiousness.

### ****Line Graphs/Charts****



* **Description:** By ploting data points on a graph and connecting them with lines to show trends over time. The x-axis represents Age of the teenagers, while the y-axis represents the number of affairs or related metrics the teenage has had.
* **Use:** To illustrate how the occurrence of teenage affairs changes over time, identifying trends, peaks, and declines.
* **At the age of 37 to 47 years old, teenagers tend to involve in more affairs, then drops completely when they turn 50 and above.**

### ****Scatter Plots****



* **Description:** This display data points on a Cartesian plane, where each point represents two variables(Age and Education). This helps in visualizing the relationship between the variables(independent and dependent variable).
* **Use:** To explore correlations between teenage affairs and external factors like education.

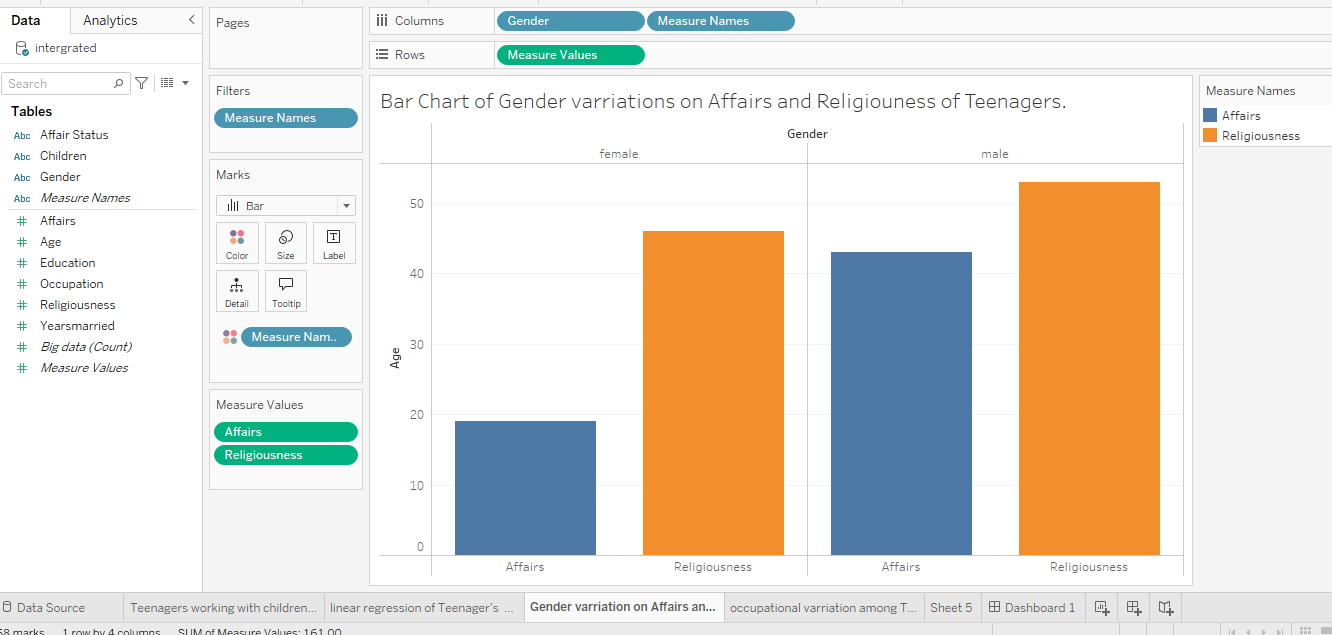
### ****Heat Maps****

* **Description:** Use color gradients to represent data values across a matrix or geographical area. Darker or more intense colors typically indicate higher values.
* **Use:** To visualize the intensity or frequency of teenage affairs in different times or locations, helping to identify hotspots or patterns.

**Insights from visualizations**

1. **Demographic Distribution (Bar Charts):**

**Insight:** This chart shows the distribution of teenage affairs across different age groups, genders, Affairs, and Religiousness.

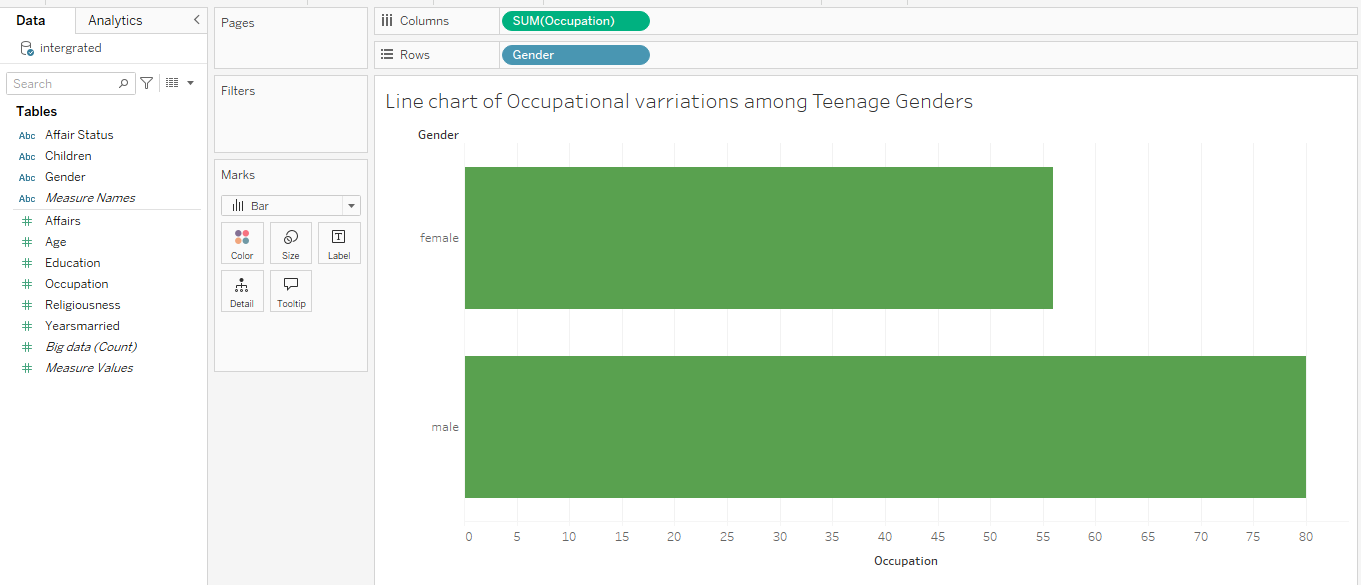


**Business Decision:** Helps target marketing efforts or tailor products/services to specific demographic segments. For instance, if data shows that teenage affairs are more common in certain regions, businesses might focus their campaigns housing areas since both male and female are having affairs and would need a house to settle.

As you can see both then to sustain their affairs above 30 years of age, implying the might have children already by that age.

1. **Trends Over Time (Line Graphs):**

**Insight:** This illustrates how the prevalence of teenage affairs has changed over time with their Occupation.

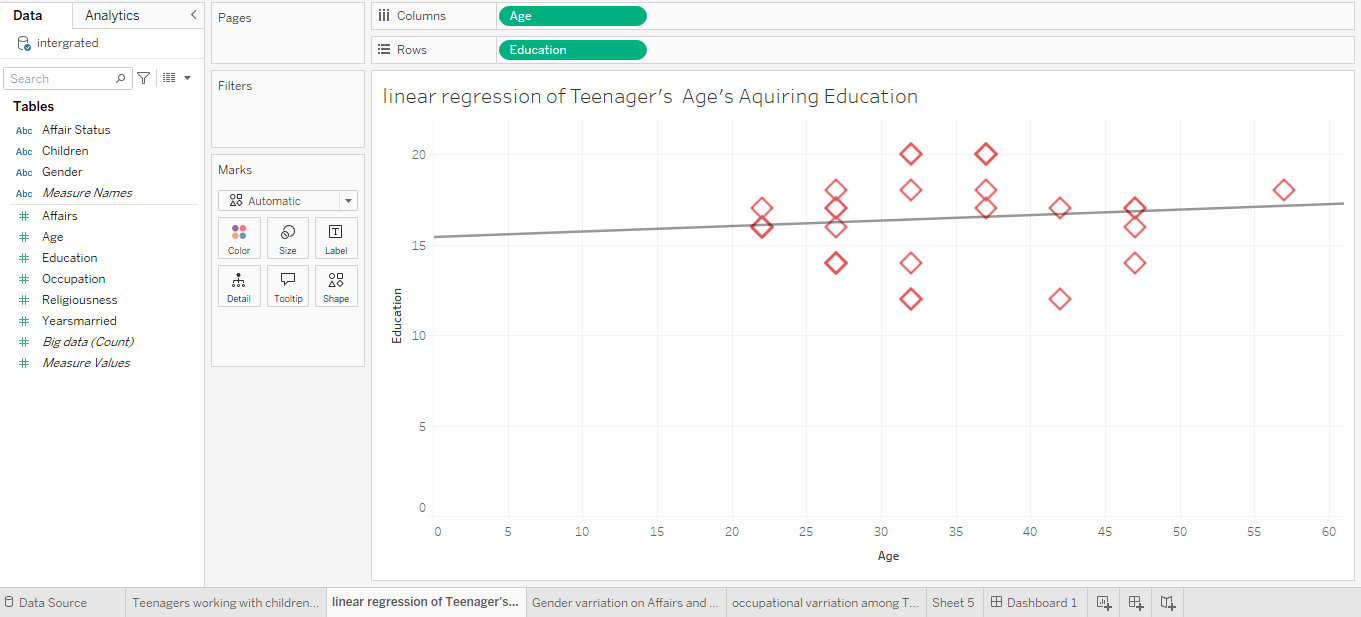


**Business Decision:** Enables businesses to anticipate future trends and adjust their strategies accordingly. For example, there’s a rising trend in Male teenage, because of new responsibilities that requires source of income verses the Female teenager who tend to drop looking for occupation, may be as a result of them being in affairs, they had a child that requires more time to look after the child, hence resigning or focusing on the Child. This gives room for newly born products in the market.

1. **Correlation with External Factors (Scatter Plots):**

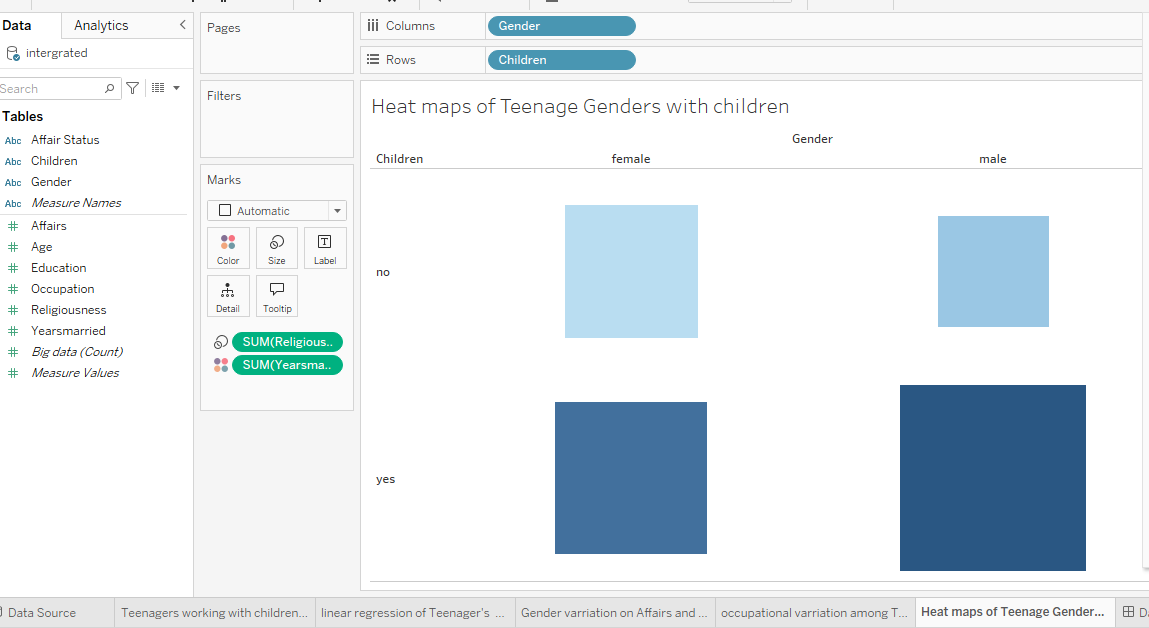
**Insight:** This shows how teenage affairs correlate with external factors of Education.

You can see that at the age of about 20 years and above, while teenager tend to starting engaging in to affair immediately after high school graduate (12) of education level



**Business Decision:** Helps in understanding the broader context affecting teenage affairs with education. There’s a strong correlation with education, schools, institutions might develop apps or content that engage teenagers more effectively so that their education performance is not affected.

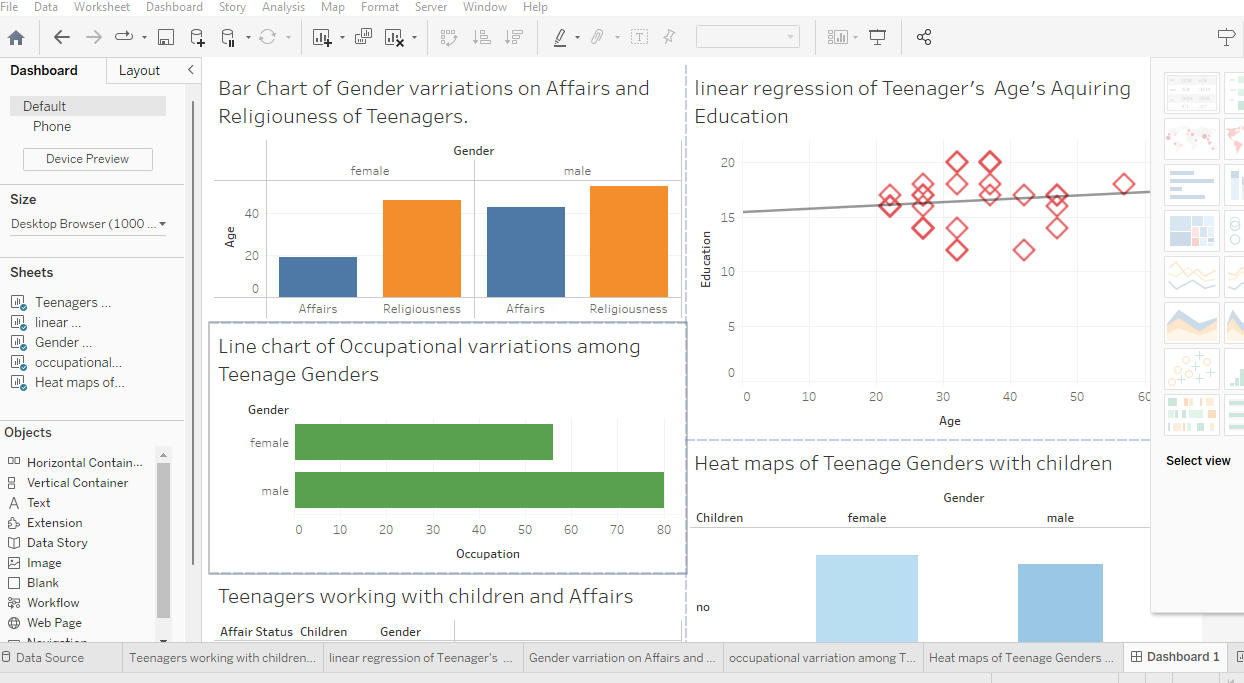
1. **Behavioral Patterns (Heat Maps):**

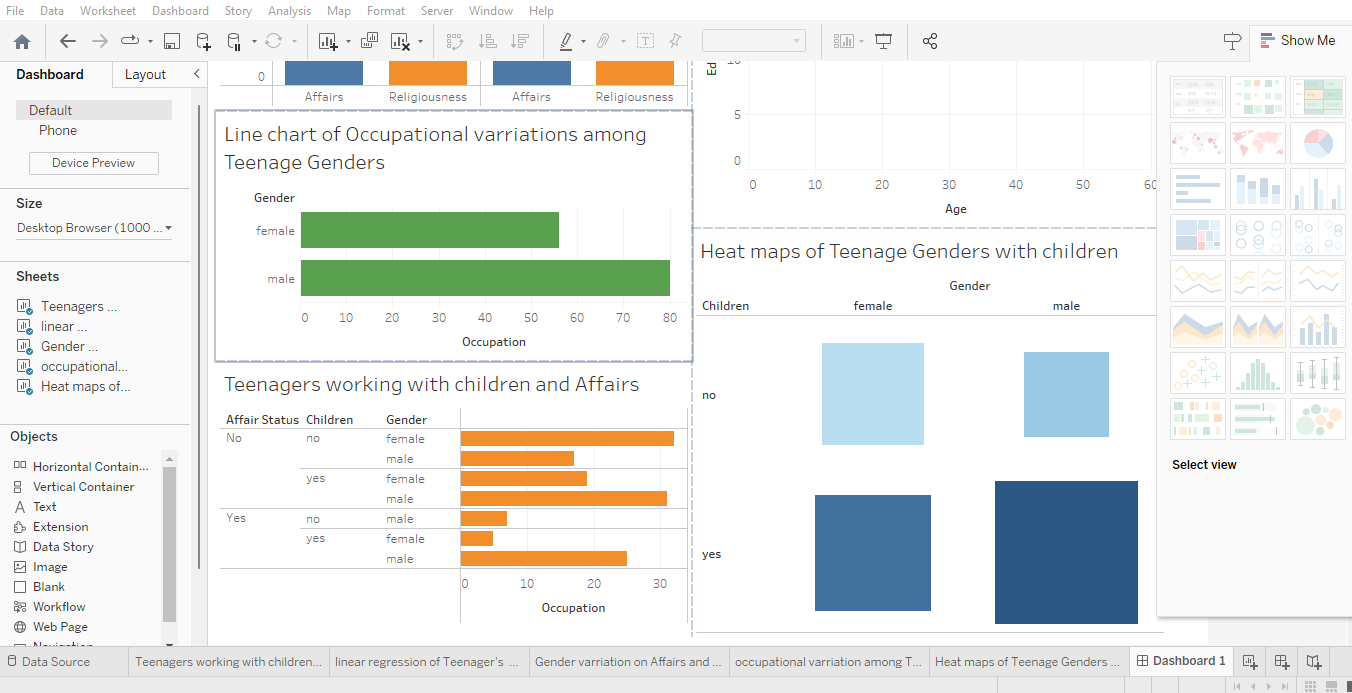


**Insight:** This displays intensity or frequency of young teenage affairs with children based on gender and having a child (yes), on not having (No).

**Business Decision:** Provides insights into peak times for engagement or potential opportunities for targeted majority of men to spend money in children shopping.

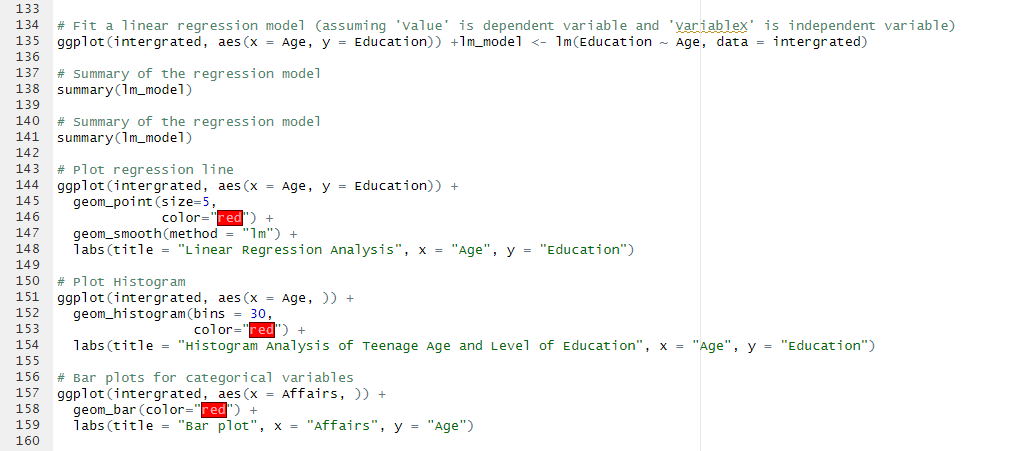
**Links/screenshots of Tableau dashboards**

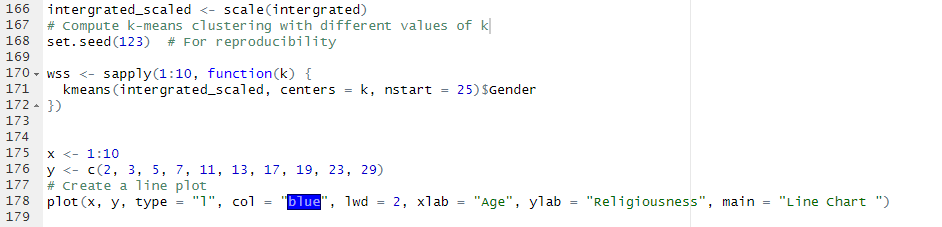
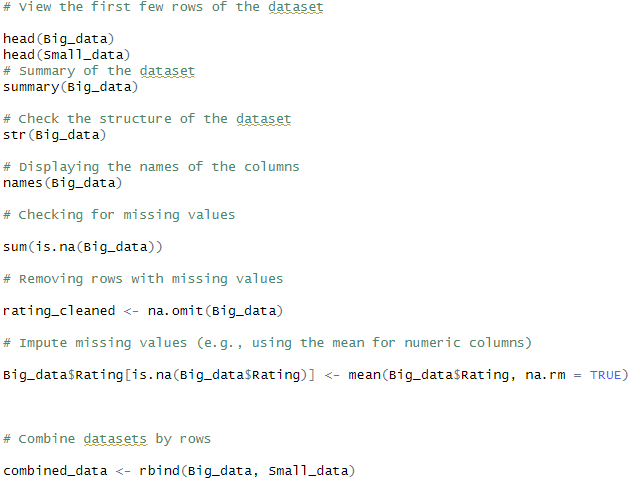
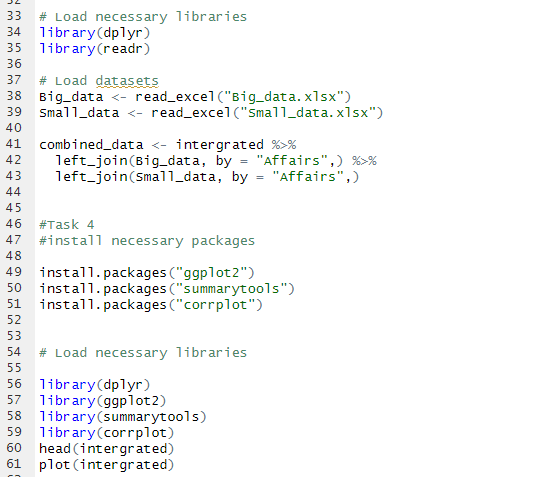
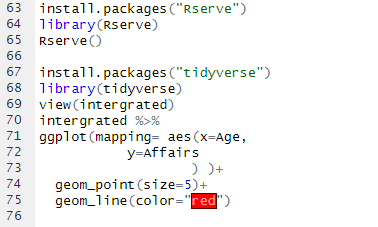
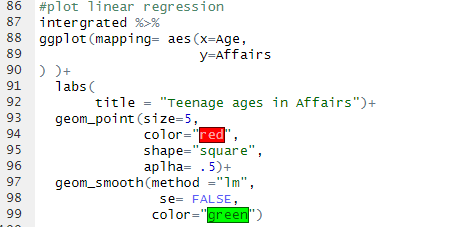
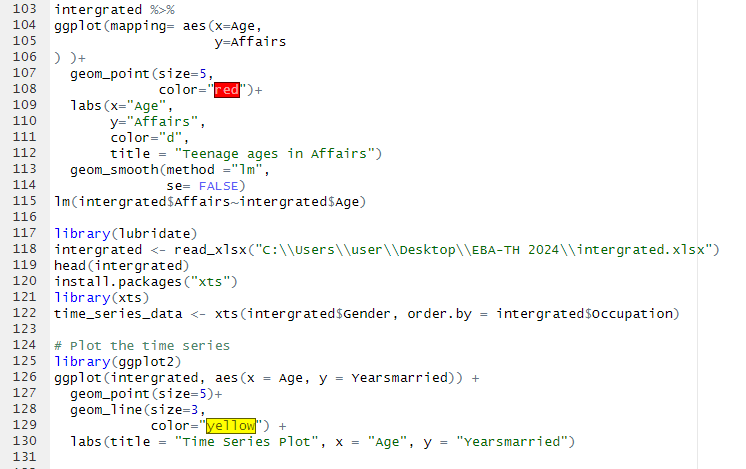




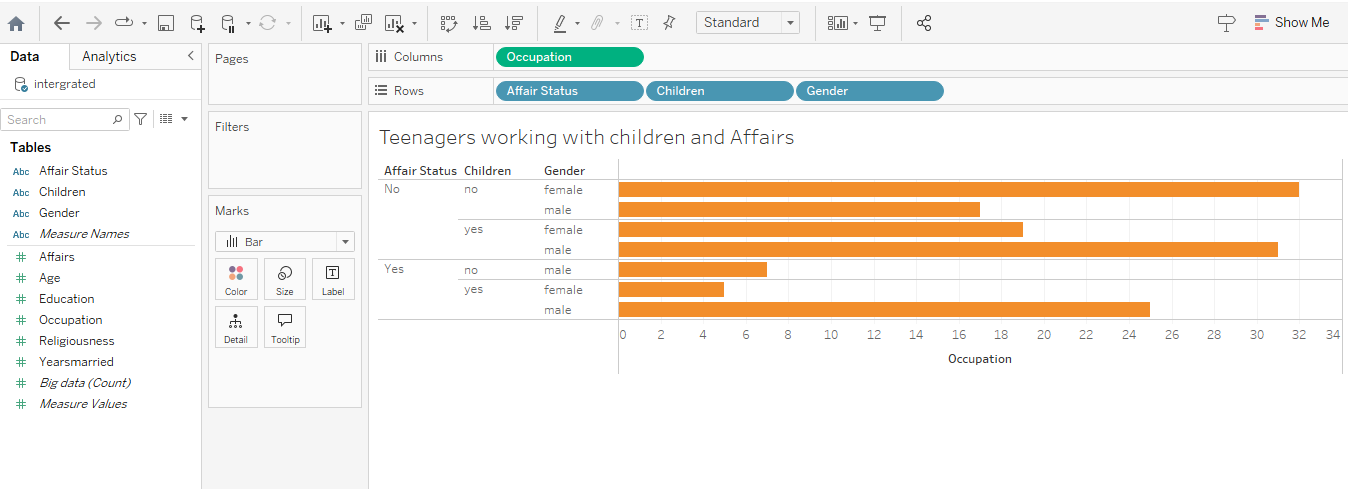
1. **Appendices**

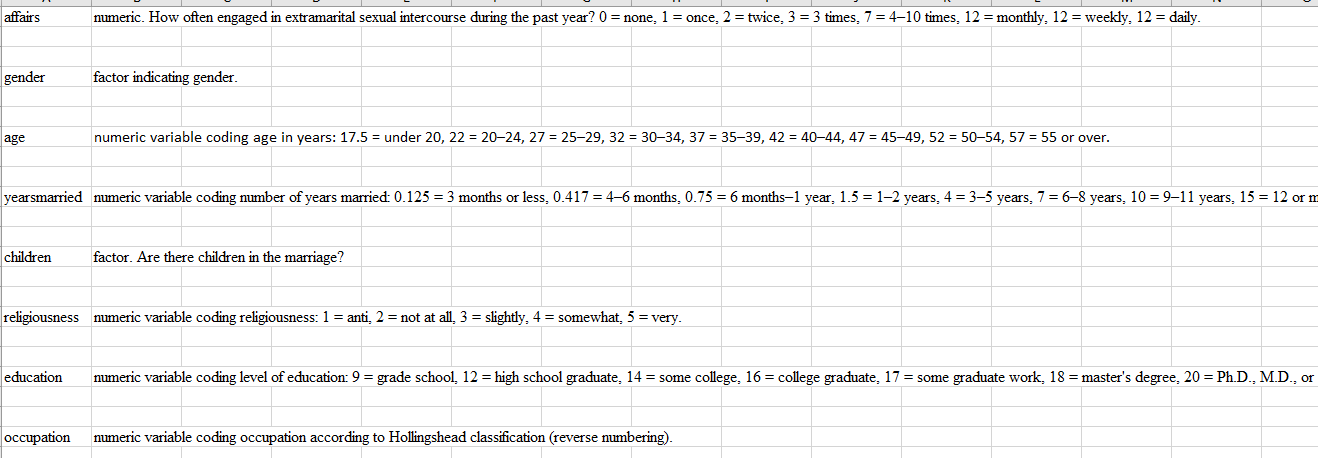
**Code snippets**





Additional charts/graphs Tools and Resources





**Data Cleaning and Analysis:**

R was used for data cleaning and analysis.

**Data Visualization:**

Tableau

**Documentation:**

Word, Excel

**Presentation:**

PowerPoint slides