

**FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES  
UITM SEREMBAN 3 CAMPUS**

**DSC651 - DATA REPRESENTATION AND REPORTING TECHNIQUES**

**INCOME CENSUS OF THE AMERICAN POPULATION**

**PREPARED BY THE MEMBERS OF GROUP 5:**

| **NAME** | **STUDENT ID** | **CLASS** |
| --- | --- | --- |
| **AZIEM HAZIQ BIN ABDUL MALEK** | **2020976839** | **N4CS2416T4** |
| **MAHFUZ BIN YUSRI** | **2020980759** | **N4CS2416T4** |
| **RAIHAN FARISS BIN RIZAL HELMY** | **2020996895** | **N4CS2416T4** |

**PREPARED FOR:  
JULIANA HAMKA BINTI KAMARODDIN**

**SUBMISSION DATE:  
12 FEBRUARY 2022**

# 

[**LIST OF TABLES**](#_koz07pjzxp0t)

[**LIST OF FIGURES**](#_9n2dcawxdodx)

[**TEAM MEMBERS**](#_e0ylu2us9gt4)

[**1.0 INTRODUCTION**](#_nu5pds5s7ni4) **1**

[**2.0 PROJECT AIMS**](#_2r1kxrr26f52) **1**

[**3.0 SCOPE OF ANALYSIS**](#_l1q4hdvjexap) **2**

[**4.0 OBJECTIVES**](#_ih9wbwr4defv) **4**

[**5.0 METHODOLOGY**](#_tclx0i8n9hz1) **5**

[5.1 Data Description](#_9zyh95oys5p2) 5

[5.2 Stages and Steps Involved](#_1kml1fwjebyf) 7

[Stage 1: Data and Context](#_ri64rdq3oogc) 7

[Stage 2: Visual Design](#_k6i2uhshk3cc) 9

[Stage 3: Storytelling Narration](#_2n3racpzvnbr) 10

[5.3 Derivative Attributes](#_99tpp7j78he) 11

[a) Age Group](#_sxvpjtq4hajq) 11

[b) Work Sector](#_1x1o8yacgd99) 12

[c) Education Level](#_4k6atsi247ss) 13

[d) Marital Class](#_vk8l29yfap1) 14

[e) Occupation Type](#_gugo9baeawu9) 16

[f) Working Hours](#_96xi2ql76ruu) 17

[g) Classification of Country Origin](#_7jb12jub3ngc) 19

[5.4 Data Issues](#_5c0us7vrbbp2) 21

[**6.0 DASHBOARD PROTOTYPE**](#_7a77kvcpf1gl) **22**

[**7.0 DATA INSIGHTS AND DISCUSSION**](#_fr0wvyan0w0) **25**

[7.1 Visual Description and Purposes](#_5og0t598ndgu) 25

[7.2 Storytelling Narration](#_djz0lcxjkmho) 27

[**8.0 CONCLUSION**](#_cm6ths9e1b6w) **37**

[**APPENDIX**](#_ope0u55rh2y2) **38**

# 

# LIST OF TABLES

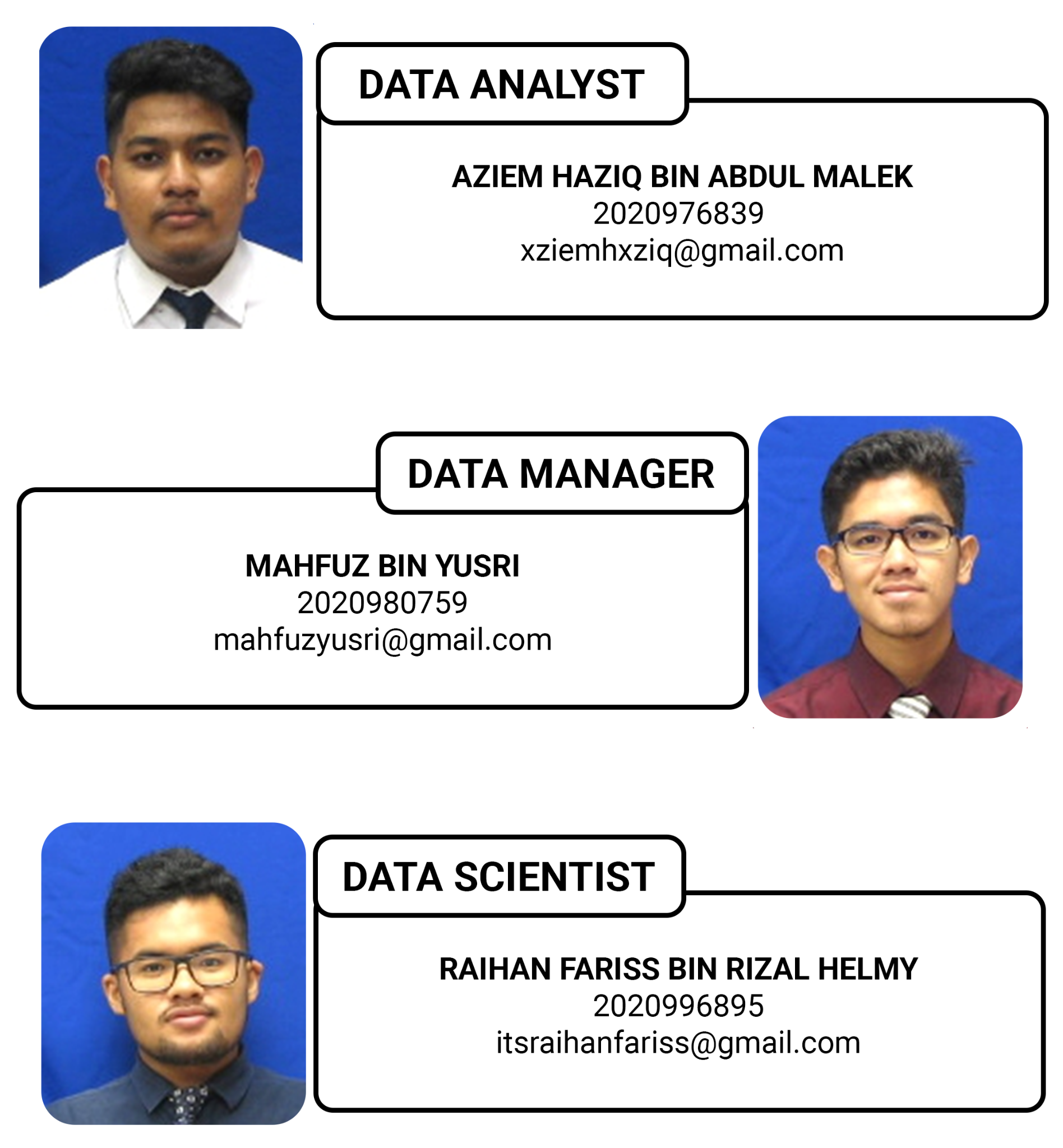
Table 1.1: Project Planning 3  
Table 5.1: Data Description of the American Population 5  
Table 5.2: Data Dictionary of the American Population 6  
Table 5.3: Gann Chart 8  
Table 5.4: Age Group 11  
Table 5.5: Work Sector 12  
Table 5.6: Education Level 13  
Table 5.7: Marital Status 15  
Table 5.8: Occupation Type 16  
Table 5.9: Working Hours 18  
Table 5.10: Country Origin 19  
Table 7.1: Descriptions and Purposes for Each Chart 25

# 

# LIST OF FIGURES

Figure 5.1: Stages and Steps Involve 7  
Figure 5.2: Age Group attributes 11  
Figure 5.3: Work Sector attributes 13  
Figure 5.4: Education Level attributes 14  
Figure 5.5: Marital Class attributes 15  
Figure 5.6: Occupation Type attributes 17  
Figure 5.7: Working Hours attributes 18  
Figure 5.8: Country Origin attributes 20  
Figure 6.1: Income Census of The American Population Dashboard 1 22  
Figure 6.2: Income Census of The American Population Dashboard 2 23

# TEAM MEMBERS



# 1.0 INTRODUCTION

The information of 32,562 adults in the United States of America (USA) from 1994 are analyzed in this research. Gender, age, race, working class, education level, income, and other factors are all included in this study's analysis. In order to develop a conclusion for this study, the trend and other key facts are noted.

The outcome of the analysis is critical for certain people, or to be more precise, the study's target audience. According to our research, by studying the feature significance of the US adult income dataset, we hope to gain a better understanding of its relevance. Between the dependent and independent variables, exploratory analysis will be performed. For our purposes, not all of the qualities are significant. The useful features will be chosen depending on the results of the various algorithms. There are numeric variables as well as numerous factors.

# 

# 2.0 PROJECT AIMS

The goal of the research is to examine the information on the income census of the population of the United States of America, as well as crucial components that should be highlighted for the upcoming research.

# 

# 3.0 SCOPE OF ANALYSIS

The project will begin on October 11th, 2021, and end on February 12th, 2022. Data gathering, data preparation, data analysis, data visualization, mock dashboard design, final dashboard design, data storytelling, and report writing are among the nine tasks completed by group members for this project. We planned tasks one by one for each week over the course of 15 weeks in order to maintain a high-quality project. As stated in Table 1, we displayed our project planning in a Gantt Chart.

Lack of knowledge, resources, procurement challenges, and timeliness might all restrict or negatively affect the success of our project. Data mistakes, inconsistencies, null values, and other difficulties obstruct data preparation. Some of the components or deliverables will not be used in our project. For example, the information on capital gain and loss is not completely reported, with the majority of respondents leaving the value at zero. After all, the major goals of our project aren't to learn more about their area or to make adjustments to it until we know and analyze the provided facts on the adults' information based on their records. There are also a few qualities that will not be employed in our project owing to their low correlation with other employee information variables. Aside from that, the scope only includes employee records from 1994, as this is sufficient for further analysis.

Several important dates are listed in this project for project milestones. First, on October 11th, the project proposal for data preparation, objectives, problem statement, and Gann chart review was submitted. This task is critical in order to move forward with the project and avoid errors. On the 10th of January, the lecturer will review the mock dashboard design. Following that, we finished the design and began dashboard narration in weeks 10 through 13. Along with the narration, we also wrote a report based on the guidelines provided. Finally, after the completion of the report in week 15, a video presentation is filmed by a group representative, and both the report and the video are submitted on February 12th.

Table 1.1: Project Planning

| **Tasks** | **W1** | **W2** | **W3** | **W4** | **W5** | **W6** | **W7** | **W8** | **W9** | **W10** | **W11** | **W12** | **W13** | **W14** | **W15** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data gathering** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data preparation** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data analysis** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data visualization** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Mock dashboard design** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Final dashboard design** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data storytelling** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Report writing** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Video Recording** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Submission (video + report)** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# 

# 4.0 OBJECTIVES

The objectives of this study are as follows:

* To determine the Level of Income based on Education Level, Work Sector and Working Hours per Week
* To determine the Work Sector based on Sex, Age and Occupation Type
* To find the relationship between Marital Status, Income Status, Education, and Country of Origin

# 

# 5.0 METHODOLOGY

## 5.1 Data Description

Table 5.1 shows the data description on the 15 attributes and Table 5.2 shows the data dictionary in this dataset.

Table 5.1: Data Description of the American Population

| **No.** | **Attributes** | **Data Description** |
| --- | --- | --- |
| **1** | age | Age of the American population, ranges from 17 up to 90 |
| **2** | workclass | Work class of the American population - Private, government or self-employed |
| **3** | fnlwgt | Final weight |
| **4** | education | Education level of the American population, which are from preschool up to doctorate. There are in total 16 levels of education in this dataset |
| **5** | education-num | The level of education, but given in numeric, 1 being the lowest (preschool) and 16 the highest (doctorate) |
| **6** | marital-status | Status of marital of the American population, which are divorced, married-AF-spouse (Armed Forces), married-civ-spouse (civilian), married-spouse-absent, never married, separated and widowed |
| **7** | occupation | Occupation of the American population, such as armed-forces, craft-repair, and farming-fishing. There are in total 15 types of occupations. |
| **8** | relationship | The role in the relationship of the American population, which are husband, not-in-family, other-relative, own-child, unmarried, wife. There are in total 6 types of relationship. |
| **9** | race | Race of the American population – white, black, Asian-pac-islander, American-Indian-Eskimo, and others. |
| **10** | sex | Sex of the American population, which are only male and female |
| **11** | capital-gain | The gain of the the American population capital |
| **12** | capital-loss | The loss of the the American population capital |
| **13** | hours-per-week | Hours per week worked by the American population, from 1 hour up to 99 hours. |
| **14** | native-country | Country of origin of population such as the United States, Cuba, and India. There are in total 42 countries of origin. |
| **15** | income | The annual income of population classified in over 50,000 USD and below or equal 50,000 USD |

Table 5.2: Data Dictionary of the American Population

| **No.** | **Attributes** | **Data Type** |
| --- | --- | --- |
| **1** | age | Integer |
| **2** | workclass | String |
| **3** | fnlwgt | Integer |
| **4** | education | String |
| **5** | education-num | Integer |
| **6** | marital-status | String |
| **7** | occupation | String |
| **8** | relationship | String |
| **9** | race | String |
| **10** | sex | String |
| **11** | capital-gain | Integer |
| **12** | capital-loss | Integer |
| **13** | hours-per-week | Integer |
| **14** | native-country | String |
| **15** | income | String |

## 5.2 Stages and Steps Involved

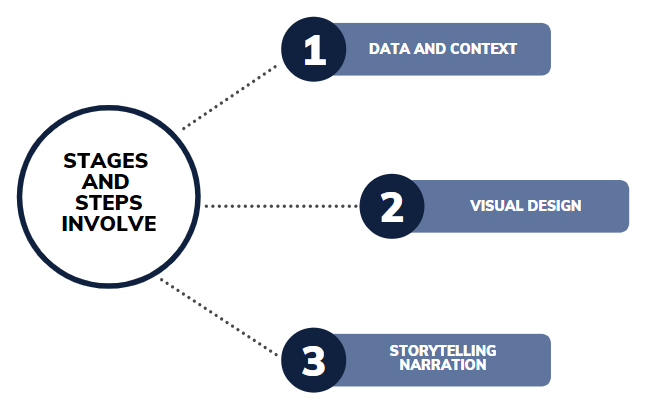


Figure 5.1: Stages and Steps Involve

### Stage 1: Data and Context

The dataset was acquired from the Kaggle website (<https://www.kaggle.com/mastmustu/income>).

This dataset consists of 15 attributes and has 32,561 records. The attributes that included in the dataset are age, work class, final weight, education, education number, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country and income. There were no missing values, duplicated data, or even inaccurate data in this data. As a result, the total number of records in this collection remains at 32,561. In order to offer relevant data information, this dataset must first conduct data description, data analysis, data classification, and data cleaning before moving on to the next stage.

Pivot table is utilized to highlight all attributes in the dataset during the data preparation phase to learn and understand more about them. Using this approach, it was identified that some of the records are very large, making it difficult to detect the chart patterns, and also, there are three attributes that are not ideal for display due to not being irrelevant to the objective of this study. The three attributes mentioned are final weight, capital loss and capital gain.

After the data has been thoroughly inspected and no further problems have been discovered, the data will be moved on to the next step, data preparation. Due to the data including some data errors such as missing values and meaningless data, it will go through data reformatting, data correction, and dataset merging during data preparation.

The target audience for this project is the employers of the United States. It is essential for employers in the United States to gain these insights to help them consider a number of essential criteria relating to the country's workforce. The relevance of the researcher's project will be discussed and assessed based on the scope and purpose of the project. Table 5.3 shows the whole progress in completing the project by using a Gann chart.

Table 5.3: Gann Chart

| **Tasks** | **W1** | **W2** | **W3** | **W4** | **W5** | **W6** | **W7** | **W8** | **W9** | **W10** | **W11** | **W12** | **W13** | **W14** | **W15** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data gathering** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data preparation** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data analysis** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data visualization** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Mock dashboard design** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Final dashboard design** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data storytelling** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Report writing** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Video Recording** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Submission (video + report)** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

### Stage 2: Visual Design

After finishing with data cleansing and preparation, the next process will proceed to data visualization. Data visualization is the process of presenting selected data in a graphical chart in order to pique the target user's interest in seeing the data shown that explain statistics and assist readers in interpreting them. At this point, the dataset will be moved to Microsoft Power BI so that the target user may obtain a clearer picture of the trends, outliers, and patterns from the various types of charts constructed by researchers.

From this dataset, nine charts were created, and all of the charts were integrated into the dashboard report. This allows people to see how the interactivity changes as they move from one chart to the next. The data can be displayed in a variety of charts, including bar charts, pie charts, line charts, and more. The charts created should be brief, well-organized, and interesting to look at. The charts in the dashboard are made with appropriate colors in mind so that they are easy to read and interpret. It also provides data labeling to help audiences or readers understand the information.

### 

### 

### Stage 3: Storytelling Narration

Storytelling narration will be written after the development dashboard at this point. It is all about effectively expressing the findings and giving the data a voice through storytelling narration. In all of the charts created for this project, trends, relationships, and outliers are recognised and examined. A good chart can tell stories by observing it, whereas a good narrative reveals the complete content of the story. The data narrative of the charts includes an introduction, a pattern and evidence as well as conclusion.

## 5.3 Derivative Attributes

### 

### a) Age Group

This attribute is created to simplify the data of age which are given in the form of numeric and giving the ages a range. This helps to paint a clearer bigger picture. The attribute taken is “age”. There are in total 5 groups of age, which are in the range as shown in Table 5.4. The result of the derivative attribute can be seen in Figure 5.2.

Table 5.4: Age Group

| **Values in attribute “age”** | **Age Group** |
| --- | --- |
| 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 | 17 - 30 |
| 31, 32, 33, 34, 35, 36, 37, 38, 39, 40 | 31 - 40 |
| 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 | 41 - 50 |
| 51, 52, 53, 54, 55, 56, 57, 58, 59, 60 | 51 - 60 |
| 60, 61, 62, 63, 64, 65, 66, 67, 68, 69,  70, 71, 72, 73, 74, 75, 76, 77, 78, 79,  80, 81, 82, 83, 84, 85, 86, 87, 88, 90 | 60+ |

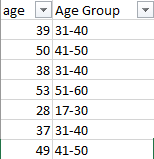


Figure 5.2: Age Group attributes

The equation used to set up the age group is as below:

=IF(A2<=30,"17-30",IF(A2<=40,"31-40",IF(A2<=50,"41-50",IF(Q2<=60,"51-60",IF(A2<=100,"60+","ERROR ")))))

### b) Work Sector

This attribute is to classify and generalize the working class of the American population. This was necessary as the values of the working class were too specific, thus hard to analyze. The attribute taken is “workclass”. There are in total 9 work classes and we generalized them into 4 work sectors as shown in Table 5.5. The result of the derivative attribute can be seen in Figure 5.3.

Table 5.5: Work Sector

| **workclass** | **Work Sector** |
| --- | --- |
| Federal-gov | Government |
| Local-gov |
| State-gov |
| Self-emp-inc | Self-Employed |
| Self-emp-not-inc |
| Never-worked | Unemployed |
| Without-pay |
| Unknown |
| Private | Private |

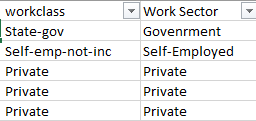


Figure 5.3: Work Sector attributes

The equation used to set up the attribute’s values are as below:

=IF(ISNUMBER(SEARCH("Federal-gov",C2)), "Govenrment", IF(ISNUMBER(SEARCH("Local-gov",C2)), "Govenrment", IF(ISNUMBER(SEARCH("State-gov",C2)), "Govenrment", IF(ISNUMBER(SEARCH("Never-worked",C2)), "Unemployed", IF(ISNUMBER(SEARCH("Private",C2)), "Private", IF(ISNUMBER(SEARCH("Without-pay",C2)), "Unemployed", IF(ISNUMBER(SEARCH("Self-emp-inc",C2)), "Self-Employed", IF(ISNUMBER(SEARCH("self-emp-not-inc",C2)), "Self-Employed", "Unemployed" ))))))))

### c) Education Level

Education Level is created to generalize the values of the education and additionally, to add a ranking. This was needed since the education was too specific and could have easily been generalized such as “11th grade”, “12th grade”, “10th grade” and “7th-8th grade” into a single class of “Highschool Grads and Lower”. The attribute taken is “education-num” because numerical data is easier to classify. There are in total 16 education numbers that were generalized into 5 education levels as shown in Table 5.6. The result of the derivative attribute can be seen in Figure 5.4.

Table 5.6: Education Level

| Values of “education-num” | Education Level |
| --- | --- |
| 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 | 1. HS Grad or Lower |
| 11, 12 | 2. Associate's Degree |
| 13 | 3. Bachelors |
| 14 | 4. Masters |
| 15, 16 | 5. Doctorate |

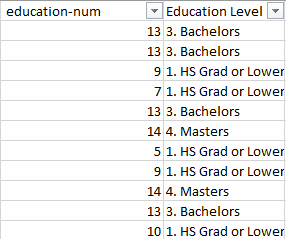


Figure 5.4: Education Level attributes

The equation used to set up the attribute’s values are as below:

=IF(G2<=10,"1. HS Grad or Lower",IF(G2<=12,"2. Associate's Degree",IF(G2<=13,"3. Bachelors",IF(G2<=14,"4. Masters","5. Doctorate"))))

### d) Marital Class

Marital Class is created to generalize the values of the attribute “marital-status”. This was needed since there are 3 types of “married”, which are married-AF-spouse (Armed Forces), married-civ-spouse (civilian) and married-spouse-absent. The value of “never-married” is also renamed as “single”. There are in total 7 marital statuses that were generalized into 5 marital classes as shown in Table 5.7. The result of the derivative attribute can be seen in Figure 5.5.

Table 5.7: Marital Status

| **Values from “marital-status”** | **Marital Class** |
| --- | --- |
| Divorced | Divorced |
| Never-married | Single |
| Married-AF-spouse | Married |
| Married-civ-spouse |
| Married-spouse-absent |
| Separated | Separated |
| Widowed | Widowed |

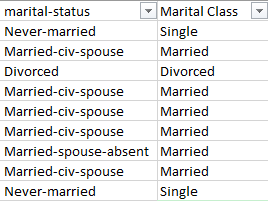


Figure 5.5: Marital Class attributes

The equation used to set up the attribute’s values are as below:

=IF(ISNUMBER(SEARCH("never-married",I2)),"Single",

IF(ISNUMBER(SEARCH("divorced",I2)),"Divorced",

IF(ISNUMBER(SEARCH("Separated",I2)),"Separated",

IF(ISNUMBER(SEARCH("widowed",I2)), "Widowed", "Married"))))

### 

### e) Occupation Type

Occupation Type is created to generalize the values of the attribute “occupation”. This was done to simplify 15 occupations of the American population into 4 general types of occupation, as shown in Table 5.8. The result of the derivative attribute can be seen in Figure 5.6.

Table 5.8: Occupation Type

| **Values in “occupation”** | **Occupation Type** |
| --- | --- |
| Adm-clerical | Employment |
| Exec-managerial |
| Handlers-cleaners |
| Sales |
| Tech-support |
| Transport-moving |
| Armed-forces | Profession |
| Machine-op-inspct |
| Prof-speciality |
| Protective-serv |
| Craft-repair | Business |
| Farming-fishing |
| Other-service |
| Priv-house-serv |
| ? | Unknown |

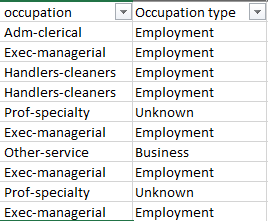


Figure 5.6: Occupation Type attributes

The equation used to set up the attribute’s values are as below:

=IF(ISNUMBER(SEARCH("adm-clerical",K2)), "Employment", IF(ISNUMBER(SEARCH("exec-managerial",K2)), "Employment", IF(ISNUMBER(SEARCH("handlers-cleaners",K2)), "Employment", IF(ISNUMBER(SEARCH("sales",K2)), "Employment", IF(ISNUMBER(SEARCH("tech-support",K2)), "Employment", IF(ISNUMBER(SEARCH("transport-moving",K2)), "Employment", IF(ISNUMBER(SEARCH("armed-forces",K2)), "Profession", IF(ISNUMBER(SEARCH("machine-op-inspct",K2)), "Profession", IF(ISNUMBER(SEARCH("prof-speciality",K2)), "Profession", IF(ISNUMBER(SEARCH("protective-serv",K2)), "Profession", IF(ISNUMBER(SEARCH("craft-repair",K2)), "Business", IF(ISNUMBER(SEARCH("farming-fishing",K2)), "Business", IF(ISNUMBER(SEARCH("other-service",K2)), "Business", IF(ISNUMBER(SEARCH("priv-house-serv",K2)), "Business",

IF(ISNUMBER(SEARCH("?",K2)), "Unknown", "error" )))))))))))))))

### 

### 

### f) Working Hours

This attribute is created to simplify the data of “hours-per-week” which are given in the form of numeric and giving the hours a range. This helps to paint a clearer bigger picture. There are in total 5 groups of working hours, which are in the range as shown in Table 5.9. The result of the derivative attribute can be seen in Figure 5.7.

Table 5.9: Working Hours

| **Values of “hours-per-week”** | **Working Hours** |
| --- | --- |
| 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 ,  11, 12, 13, 14, 15, 16, 17, 18, 19, 20 | 1 - 20 |
| 21, 22, 23, 24, 25, 26, 27, 28, 29, 30,  31, 32, 33, 34, 35, 36, 37, 38 ,39, 40 | 21 - 40 |
| 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,  51, 52, 53, 54, 55, 56, 57, 58, 59, 60 | 41 - 60 |
| 61, 62, 63, 64, 65, 66, 67, 68, 70,  72, 73, 74, 75, 76, 77, 78, 80 | 61 - 80 |
| 81, 82, 84, 85, 86, 87, 88, 90,  91, 92, 94, 95, 96, 97, 98, 99 | 81 - 100 |

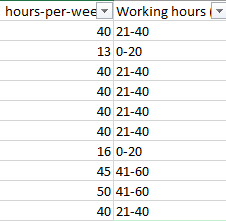


Figure 5.7: Working Hours attributes

The equation used to set up the attribute’s values are as below:

=IF(R2<=20,"0-20",IF(R2<=40,"21-40",IF(R2<=60,"41-60",IF(R2<=80,"61-80",IF(R2<=100,"81-100","ERROR ")))))

### g) Classification of Country Origin

Since the data is regarding the American population, it is expected that the majority of the populations are from a native background. However, in this dataset, the number of native citizens outnumber the non-native citizen by a landslide, which is 29,170 out of 32,561, leaving the non-natives not only in the very minority, but also very minimal since there are a total of 41 types of non-natives. Therefore, to simplify the data, we generalized respondents that the value of “native-country” is “United-States” as “Native” in our derived attribute “Country Origin” and the other values other than “United States” as “Non-Native”, as shown in Table 5.10. The result of the derivative attribute can be seen in Figure 5.8.

Table 5.10: Country Origin

| **Values of “native-country”** | **Country Origin** |
| --- | --- |
| United-States | Native |
| Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands | Non-Native |

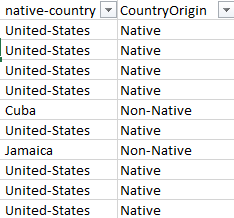


Figure 5.8: Country Origin attributes

The equation used to set up the attribute’s values are as below:

=IF(ISNUMBER(SEARCH("United-States",T2)), "Native", "Non-Native")

## 

## 5.4 Data Issues

There are three (3) main issues in this dataset across four (4) variables. The first issue being that an attribute that seems to be very useful, usable and important during data search is rendered useless since the definition of the attribute is not clear. Said attribute is “fnlwgt”, which means “final weight”. Usually in a dataset, weight is used to re-balance a dataset in order to more accurately reflect the population or include a multiplier that projects the result of a larger universe. However during data analysis it was found out that there is no definition as to what “fnlwgt'' is supposed to be used for and how to use it, thus making it impossible to use. The second issue is that the values of the attribute “capital-gain” and “capital-loss” are too scarce thus making it also unusable. The last issue is that the most important attribute, “income”, is already classified into two classes of more than 50,000 USD annual or less than or equal 50,000 USD annual. The problem occurs because the overgeneralization of data hinders this research from making a more in-depth evaluation of the income status of the American population. Thus, the results obtained in this paper may not represent the actual situation in the American population.

# 6.0 DASHBOARD PROTOTYPE

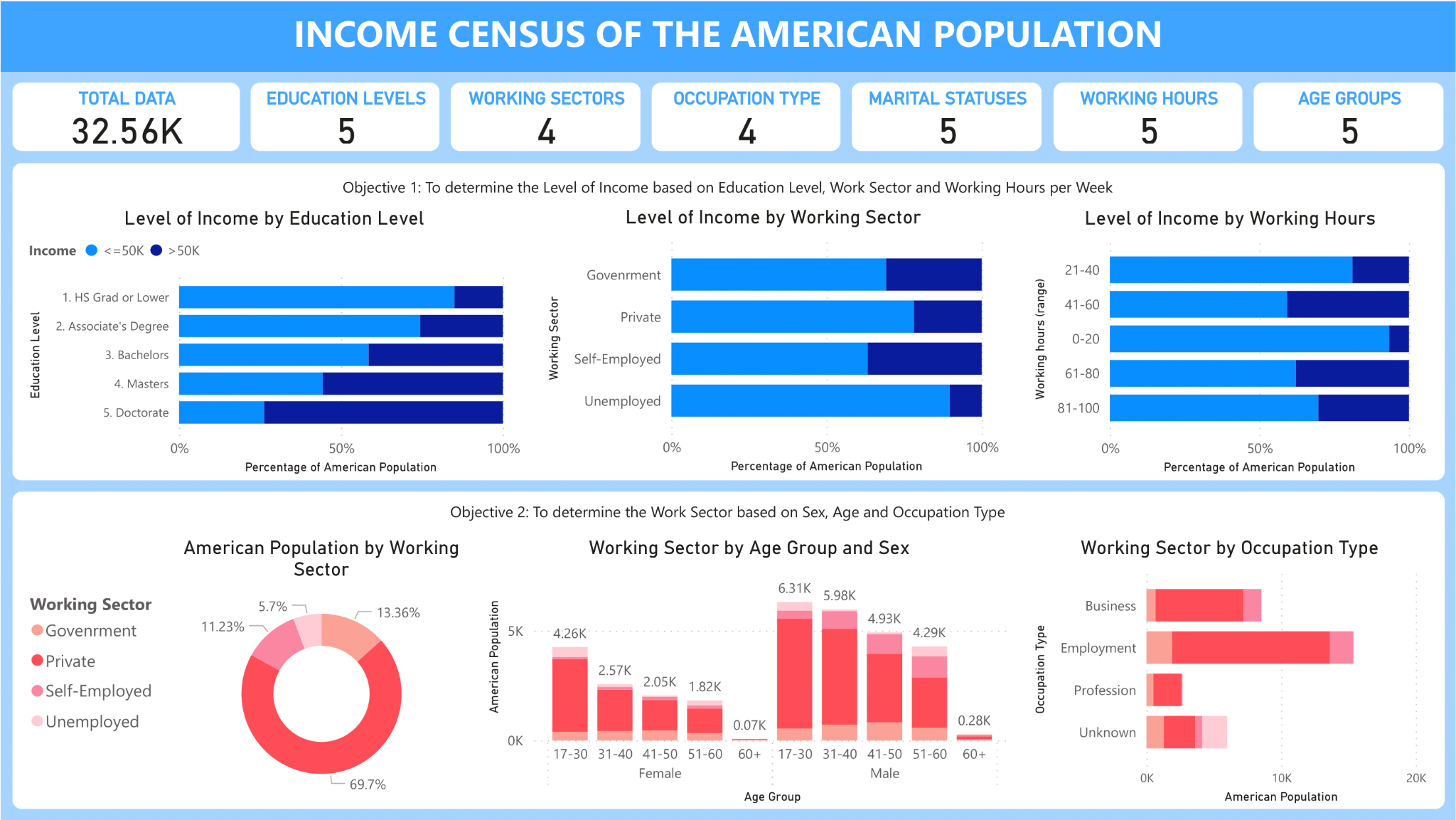


Figure 6.1: Income Census of The American Population Dashboard 1

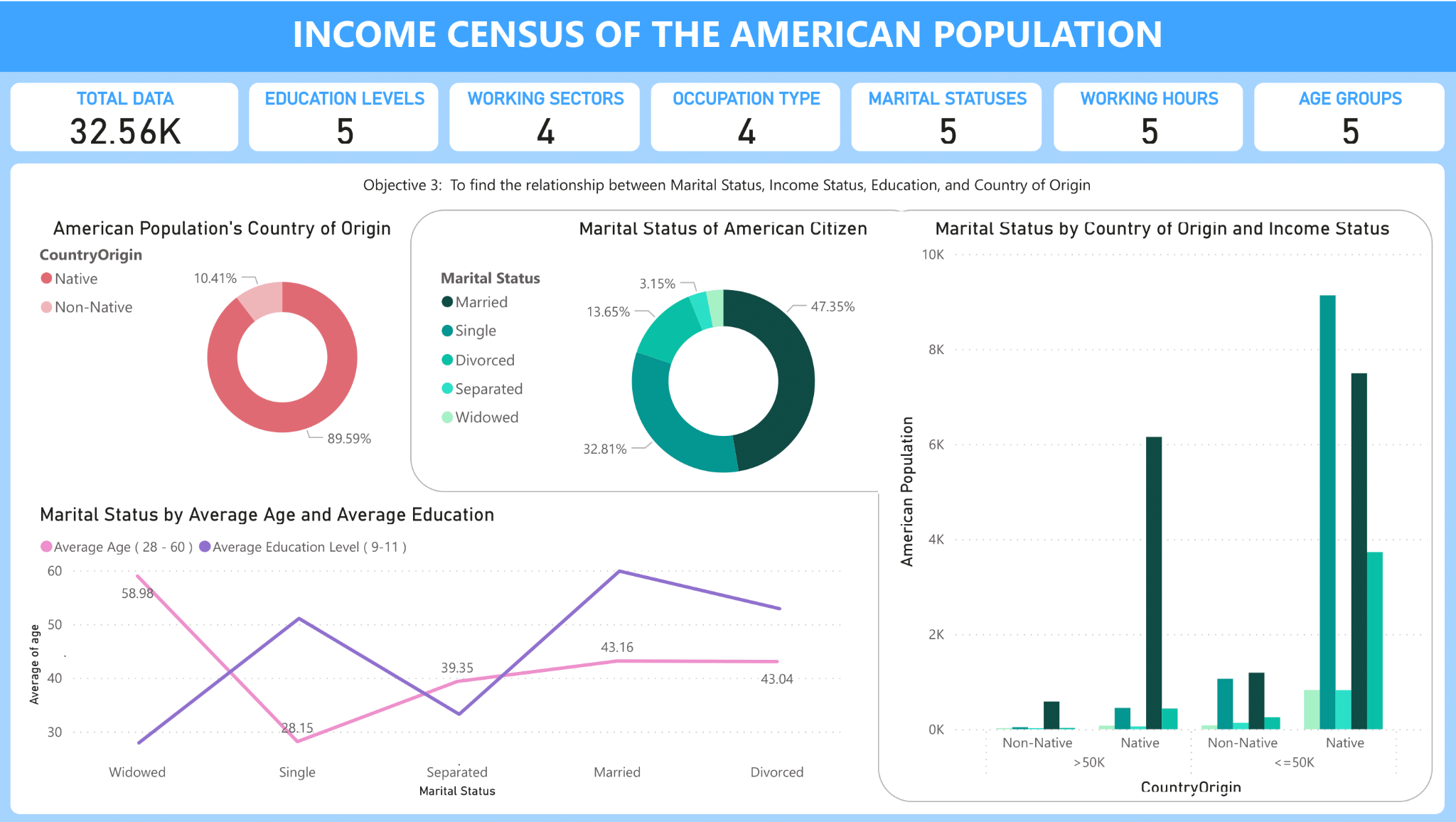


Figure 6.2: Income Census of The American Population Dashboard

Microsoft Power BI was used to visualize the data. Data from a census is used to create the graphic, which depicts the American population. To be more precise, this data visualization is aimed at the general public and illustrates a variety of statistics and relationships relating to the general population of the United States. Our target audience falls under the generalist group. Because our audience comes from a variety of educational levels, we made sure that our data visualization was simple to grasp. The dashboard is used for data analysis.

The dashboard is made up of various components. To begin, the dashboard displays five main value cards: total data, educational levels, working sectors, employment type, marital statuses, working hours, and age groupings. The dashboard's title is "Income Census of the American Population." Because the dataset's information originates from the American income census, the dashboard's title was chosen accordingly. The term was also kept broad since the data is intended to offer a broad perspective of demographic changes in the United States.The dashboard is made up of ten distinct types of charts that depict the data and its tendencies. In general, the goal of this dashboard is to reveal hidden important data for the general audience to assess based on qualities.

The colors picked for the dashboard are according to two purposes, which are decoration and representation. The colors blue and white are the most popular for ornamentation, as seen by the chart's borders and the use of borders. Blue was chosen since it has been scientifically shown that blue relaxes the mind, which is why many social media hues, such as Twitter and Facebook, are blue. For the representation, there are five main dominant colors which are blue, red, green, pink, and purple. The dominant colors differentiate each category such as blue for level of income, red for working sector, green for marital status, pink for nativity, and purple for age.

Furthermore, the dashboards are read from left to right, and each row is separated to represent one objective. For example, in the first dashboard, the viewer begins reading from the top section, which is the first objective, and starts with the furthest left chart, "Level of Income by Education Level," before moving to the right chart next. The second portion below, which relates to the second objective, is the same, as does the second page. After the dashboards have been finalized, they are published on the Power BI website in which the corresponding link is <https://rb.gy/94zcqp>.

# 7.0 DATA INSIGHTS AND DISCUSSION

## 7.1 Visual Description and Purposes

This chapter discusses the descriptions and purposes for all the charts in the dashboard.

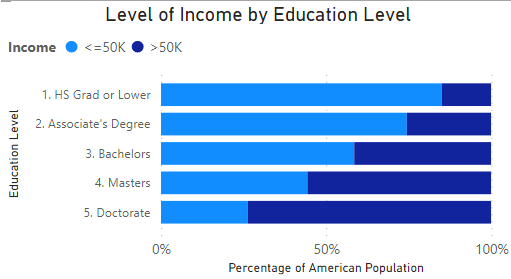
Table 7.1: Descriptions and Purposes for Each Chart

| **No** | **Title of Chart** | **Description** | **Purpose** |
| --- | --- | --- | --- |
| 1 | Level of Income by Education Level | Income and education level attributes were used in the 100% stacked bar chart. | To determine the pattern of level of income for each education level. |
| 2 | Level of Income by Working Sector | Income and work sector attributes were used in the 100% stacked bar chart. | To determine the pattern of level of income for each working sector. |
| 3 | Level of income by Working Hours | Income and working hours attributes were used in the 100% stacked bar chart. | To determine the pattern of level of income based on the working hours. |
| 4 | American Population by Working Sector | Working sector attributes were used in the donut chart. | To display the American population in each working sector. |
| 5 | Working Sector by Age Group and Sex | Working sector, sex and age group attributes were used in the stacked column chart. | To find the patterns of the working sector in America based on sex and age group. |
| 6 | Working Sector by Occupation Type | Working sector and occupation attributes were used in the stacked bar chart. | To find the patterns of the working sector in America based on the type of occupation. |
| 7 | American Population’s Country of Origin | Country origin attributes were used in the donut chart. | To display the country of origin in the American population. |
| 8 | Marital Status by Average Age and Average Education | Marital status, age group and education level attributes were used in the line chart. | To find the relationship between marital status of Americans and the average age and education. |
| 9 | Marital Status of American Citizen | Marital status attributes were used in the donut chart. | To discover the number of Americans for each marital status. |
| 10 | Marital Status by Country of Origin and Income Status | Marital status, country origin and income attributes were used in the clustered column chart. | To compare the American population for each marital status based on their origin country and income status. |

## 

## 7.2 Storytelling Narration

**Chart 1: Level of Income by Education Level**

****

The first chart in the dashboard is Level of Income by Education Level. This is a 100% stacked bar chart, which represents the percentage of the American population which have annual income higher than 50,000 USD or less than or equal 50,000 USD according to their level of education. The x-axis represents the level of percentage of level of annual income, where light blue is equal or less than 50,000 USD, or dark blue which is more than 50,000 USD. The y-axis represents the level of education, which is ranked from 1 to 5, 1 being the lowest level of education - Highschool (HS) graduates or lower, and 5 being the highest, doctorates. Chart 1 shows a perfectly representable pattern, where the higher the level of education, the more the percentage of the American population that has the higher annual salary increases. The population who are HS graduates and lower have the lowest percentage of an annual salary exceeding 50,000 USD by only 15%. People with an associate’s degree have a slight increase with a percentage of 25.56%. Bachelor holders on the other hand have a moderate increase in percentage of 41.48% compared to Associate’s Degree holders. Repeating the trend, Master’s holders also have an increased percentage of 55.66%, while finally Doctorates holders have the highest percentage of 73.71% in having annual salaries exceeding 50,000 USD. In conclusion, the population with a higher education level tend to have an increased chance for having a higher salary.

**Chart 2: Level of Income by Working Sector**

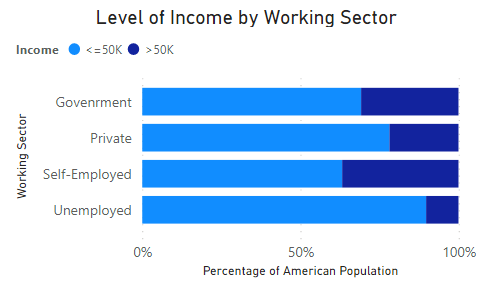
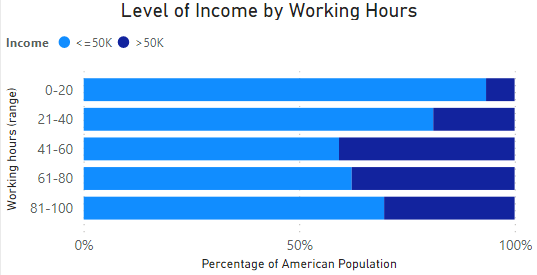
****

Chart 2 is also a 100% stacked bar chart regarding the Level of Income by Working Sector, which represents the percentage of the American population which have annual income higher than 50,000 USD or less than or equal 50,000 USD according to their working sector. From the chart, the population with the highest rate of achieving more than 50,000 USD annual salary are of the population that are self-employed or working in the government sector, having 36.82% and 30.82% respectively. The population that work in the private sector don’t fall far behind, having 21.87% of the population making more than 50,000 USD annually. As a conclusion, the population that works in the government sector or are self-employed tend to receive more salary than people that work in the private sector. Additionally, working in the government sector also gives the benefit of extra remuneration and benefits that should not be overlooked.

**Chart 3: Level of Income by Working Hours**

****

The third chart is again - a stacked bar chart - which shows the percentage of the American population which have annual income higher than 50,000 USD or less than or equal 50,000 USD according to their range of working hours per week. The y-axis represents the hours worked by the population in 5 levels of range, ranked in an increasing order. From observation, we can generally divide the working hours into two, one being 0-20 and 21-40 hours, and group two being 41-60, 61-80, and 81-100 hours. Group two surpasses group one by a lot in terms of the percentage of people that earn more than 50,000 USD annually. For the percentage of the population that earns more than 50,000 USD annually, the members of group one, being 0-20 and 21-40 hours per week, have the percentage of 6.66% and 18.9% respectively. Whereas on the other hand, for group two, consisting of the range 41-60, 61-80, and 81-100 hours per week have the percentage of 40.75%, 37.8%, and 30.29% respectively. in conclusion, we can observe that higher working hours does not necessarily guarantee a higher amount of earnings. It is important to be able to self-judge how much one’s working hour is worth to be able to maximize their earnings.

**Chart 4: American Population by Working Sector**

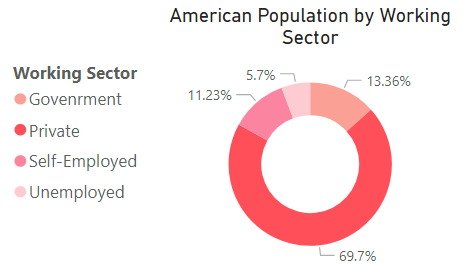
****

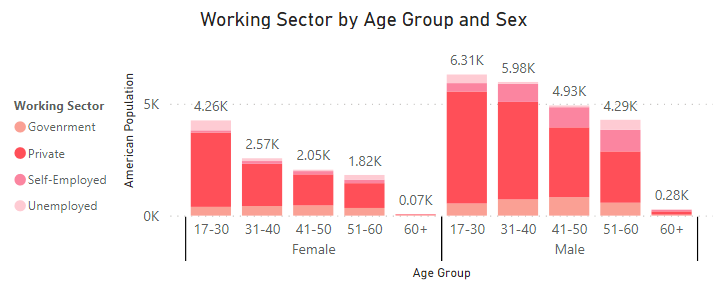
Chart 4 is a donut chart that shows the American population by working sector. It displays the percentage of Americans in all four work sectors in the United States. The four work sectors mentioned are government, private, self-employed, and unemployed.

Based on the donut chart, most of the Americans in the United States are working in the private sectors. Meanwhile, the unemployed sector has the lowest number of American population compared to the other working sectors.

As evidence, the private sector has the highest percentage value of 69.7%. Following the private sector as the highest percentage, government and self-employed sectors are in the second and third rank with the percentage of 13.36% and 11.23% respectively. Furthermore, the working sector that has the lowest percentage of 5.7% is the unemployed sector.

To conclude, a massive portion of the American population works within the private sector in the United States. This shows that the occupation in the private sector in the United States may have been more developed compared to the other work sectors. Furthermore, the Americans may gain more benefits when working in the private sectors.

**Chart 5: Working Sector by Age Group and Sex**

****

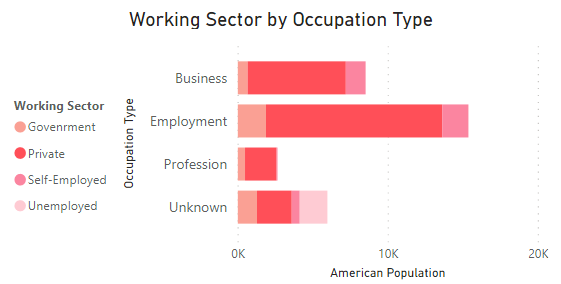
The fifth chapter is a stacked column chart that displays the working sector by age group and sex. It shows the number of Americans in each work sector according to the age group and sex.

Based on the stacked column chart, most of the Americans in the United States are male with the age from 17 to 30 years old working in the private sectors. Moreover, the chart also shows that there are more male population compared to females in the working sectors of the United States. In addition, observation can be made from the chart for both genders that when the age group increases, the American population in the working sectors decrease.

As proof, the American population of male from the age of 17 to 30 years old working in the private sector has the highest number of 6.31k people. Furthermore, the highest female workers are from the age group from 17 to 30 years old with the value of 4.26k people. Yet, it does not surpass the fourth highest male workers (age group from 51 to 60 years old) that has the value of 4.29k. Additionally, it also shows that stacked column charts begin to drop when the age group is increased.

As a conclusion, a lot of youth in America at the age of 17 to 30 years old actively seek a job especially in the private sector. Meanwhile, most American started to retire at the age of 60 and above. The factors of the retirement could include the loss of their energy and strength, deteriorate health, and so on.

**Chart 6: Working Sector by Occupation Type**

****

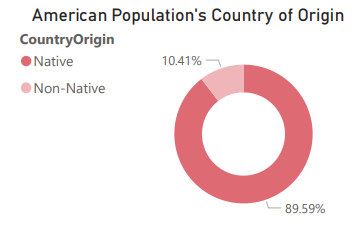
Chapter 6 is a stacked bar chart that shows the working sector by occupation type. There are four types of occupation in the dataset, which is business, employment, profession, and unknown.

According to the chart, the private sector is still the highest in terms of the American population compared to the other sector. In addition, most Americans work in the employment occupation in the private sector of the United States. Besides that, the number of Americans who are unemployed are only in the unknown occupation type.

As evidence, the employment occupation has a total number of 11,741 Americans who work in the private sector. Furthermore, there are no unemployed Americans in all types of occupation except in the unknown occupation with the value of 1,843 people.

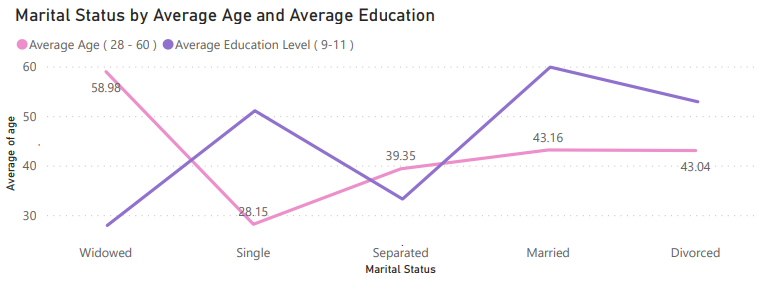
To conclude, most of the Americans in the United States are interested to work in the employment occupation type in the private sector. Moreover, it is a positive sign for the Americans that the United States has no records of unemployed people in three out of four occupation types.

**Chart 7 : American Population's Country of Origin**

****

This donut chart shows the American population’s country of origin which is based on two factors which are native and non-native. The highest number of people are in the group of native compared to non-native. It showed that the majority of people in the United State of America are native Americans. Based on the previous trend, the percentage of native americans is 89.59% or equal to 29170 peoples. Then, the percentage of non-native Americans is 10.41% or equal to 3391. In conclusion, the majority of the population is native americans while the remaining of them is non-native americans.

**Chart 8 : Marital Status by Average Age and Average Education**



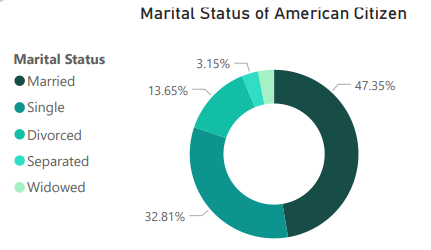
The eighth chart in the dashboard is Marital Status by Average Age and Average Education. This line chart represents the marital status of American citizens based on average age and average education. The marital status axis above shows 5 status which are Widowed, Single, Separated, Married, and Divorced. The average of the age axis above shows 4 values which are 30, 40, 50, and 60. Then, there is also a line that represents the average education level rate from 9 to 11. Based on the chart, the difference between the average age is decreasing across all marital status. The highest average age is from widowed with 58 years old , while the lowest one are from single with 28 years old. The highest average education level is from married status with 10.29, while the lowest one is from widowed with 9.09. In conclusion, the higher the education level the higher the average age. Hence, married citizens proved that a high education level is not an exception and can be successful.

# 

# 

# 

**Chart 9 : Marital Status of American Citizen**

****

This donut chart shows the marital status of American citizens with the marital status based on married, single, divorced, separated, and widowed. Based on the chart, married citizens have the highest number of citizens compared to other marital status with widows having the lowest number of citizens. It showed that the majority of people in the United State of America are married. Based on the previous trend, the percentage of married citizens is 47.35% or equal to 15417 peoples. Then, the percentage of widows is 3.05% or equal to 993. In conclusion, the majority of the population are married and dominating other marital status.

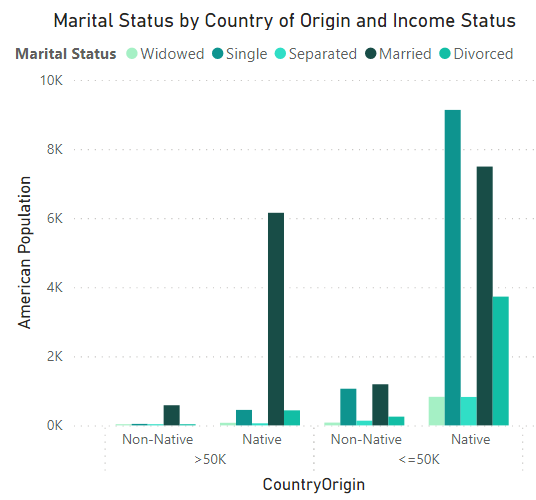
# 

# 

# 

# 

**Chart 10 : Marital Status by Country of Origin and Income Status**

****

The tenth chart in the dashboard is Marital Status by Country of Origin and Income Status. This is a column chart; it represents the American population in the country according to their origin, income status and marital status. The region axis shown above has 2 origins which are native and non-native while income status has 2 categories which is more than 50k and less or equal to 50k and marital status divided into 5 categories, married, single, separated, divorced and widowed. Based on the chart, there is an increasing pattern for native citizens. It means the highest number of citizens are from native citizens with single status and have an income of less or equal than 50k of 9131 citizens. This is followed by native citizens with married status and have an income of less or equal than 50k of 7491 citizens. The non-native citizens have the lowest number of citizens therefore they are the minority in the country. To conclude, native Americans have been mostly dominated by single and married citizens for both income status.

# 8.0 CONCLUSION

# 

Finally, we derive a conclusion from the analysis of the charts in a dashboard, hence a narration of overall conclusion has been made. This includes the explanation on the most significant trend, relationship and outliers. We have provided the evidence including the chart name.

It is apparent that this data visualization approach has a wide range of applications. All of the project's goals were met, and it was a huge success. Some data visualization was shown in the previous chapter, and the data pattern was explored. As can be seen, there are various phases to creating an efficient data visualization that makes the data simple to understand and leaves a lasting impression on the audience. Additionally, a few new properties have been established in order to simplify data visualization in a more understandable pattern.

# 

# 

# 

# 

# 

# 