**FINAL PROJECT**

**MILESTONE - II**

**DESCRIPTIVE STATISTICS**

**GROUP 8**

**Team Members List:**

Varshitha Inaganti(11637772)

Chathurya Thimmapuram(11479590)

Shashank Jha(11634967)

Sujith Kumar Jonnadula(11599424)

**TABLE OF CONTENTS:**

1. INTRODUCTION………………………………………………………..………..… 3
2. LIST OF VARIABLES……………………………………………………………… 4
3. SUMMARY OF DESCRIPTIVE STATISTICS……………………………………... 4
4. DESCRIPTIVE STATISTICS OF CATEGORICAL VARIABLES…………………..6

4.1 FREQUENCY DISTRIBUTION OF PERSONALITY CATEGORICAL VARIABLES

4.2 FREQUENCY DISTRIBUTION OF PERSONALITY, BUSINESS AND FEMALE VARIABLES

4.3 PERCENT FREQUENCY DISTRIBUTION OF FEMALE OF SALES REPRESENTATIVES

4.4 PERCENT FREQUENCY DISTRIBUTION OF COLLEGE OF SALES REPRESENTATIVES

1. DESCRIPTIVE STATISTICS OF QUANTITATIVE VARIABLES…………………10

5.1 FREQUENCY DISTRIBUTION OF CUSTOMER SATISFACTION SCORE OF DISCRETE VARIABLE AND AGE, FEEDBACK OF CONTINUOUS VARIABLES

5.2 FREQUENCY DISTRIBUTION OF PERSONALITY BY AGE

1. SIMPSON’S PARADOX…………………………………………………………… 17
2. CONCLUSION…………………………………………………………………….. 18

**1. Introduction**

The collection includes useful data that is relevant to the Tech sales sector. The dataset offers details on a number of sales reps' performance, traits, and salary as well as other relevant information. The dataset includes details on the business focus, age, gender, number of years of experience, education level, personality type, certifications, feedback received, compensation, and customer satisfaction rating of the sales representative.

The mentioned industry relies heavily on data analysis because it offers vital information for assessing sales performance, predicting trends, comprehending client behavior, enhancing sales methods, and maintaining competitiveness. Businesses may decide wisely on promotions, rewards, and resource allocation by looking at data on sales representative performance, customer satisfaction ratings, and market trends. Furthermore, by identifying client preferences and customizing products and services accordingly, data analysis enables businesses to stay ahead of the competition. Overall, data analysis equips companies to increase sales and accomplish their goals in a market that is changing quickly.

**2. List of variables:**

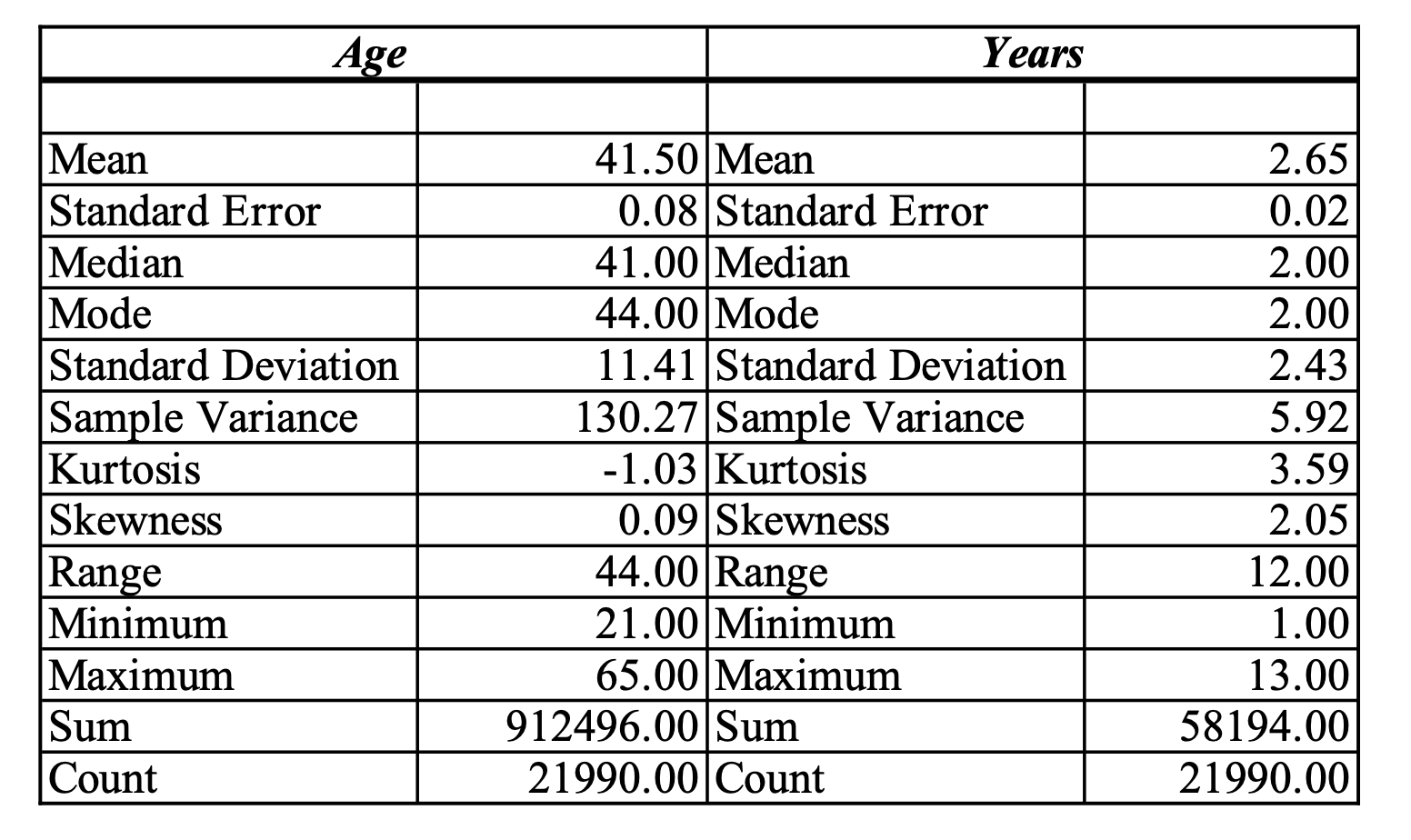
The qualitative and quantitative groupings of variables in data analysis are the two basic categories. Although they cannot be mathematically operated on, qualitative variables reflect traits or categories. Gender, college and personality are a few examples of qualitative factors. Alternatively, quantitative variables are made up of numerical values that can be measured and quantitatively examined. Age, height, and length are some illustrations of quantitative variables.

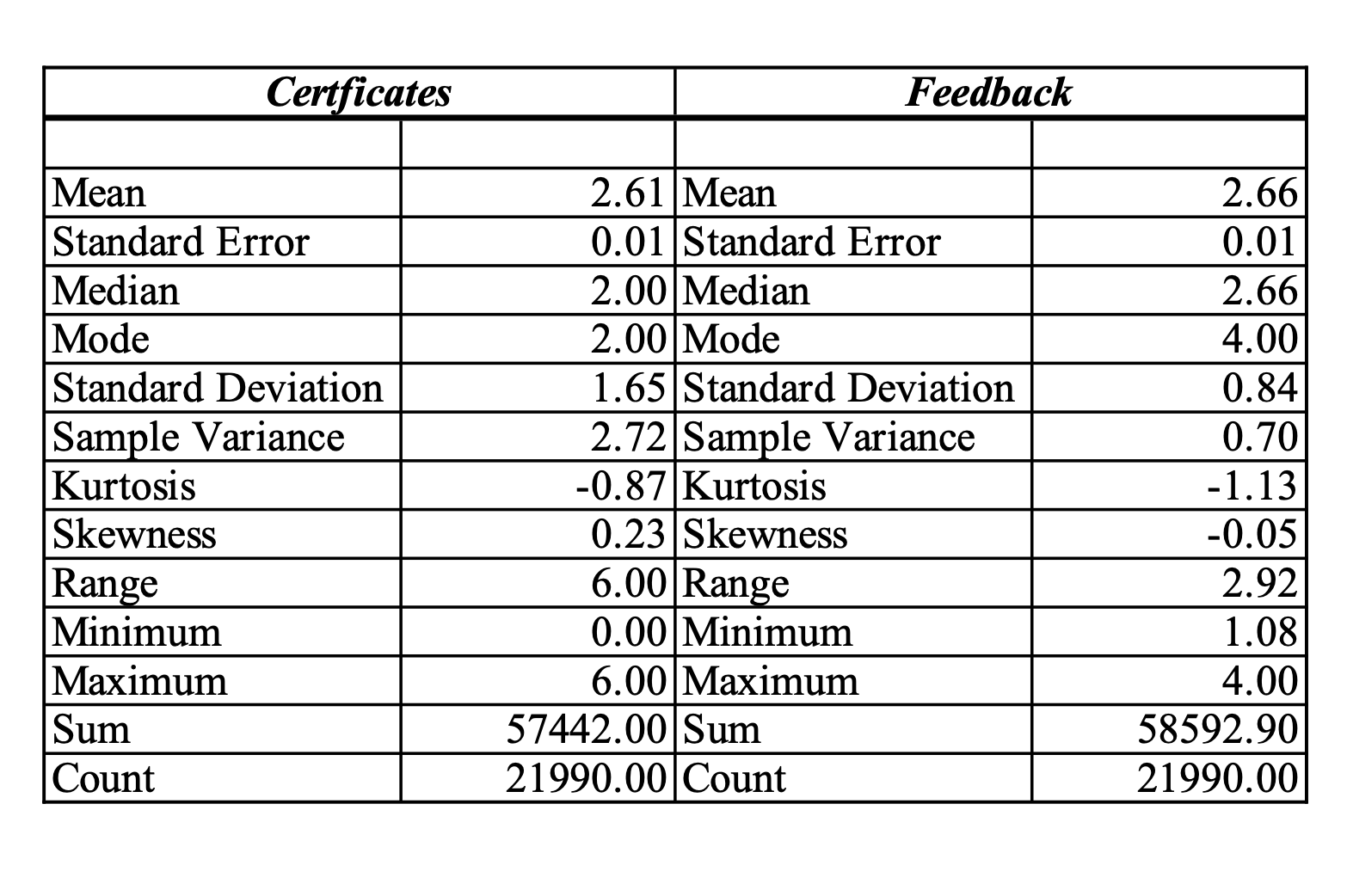
There are two further classifications of quantitative variables: discrete and continuous variables. Discrete variables can only take on a limited or countable amount of values. Consider discrete variables as "dogs" that can give birth to 1, 2, or 10 pups, depending on the value. The variable's acceptable range of values is defined and constrained. Continuous variables, on the other hand, are capable of an endless variety of values.

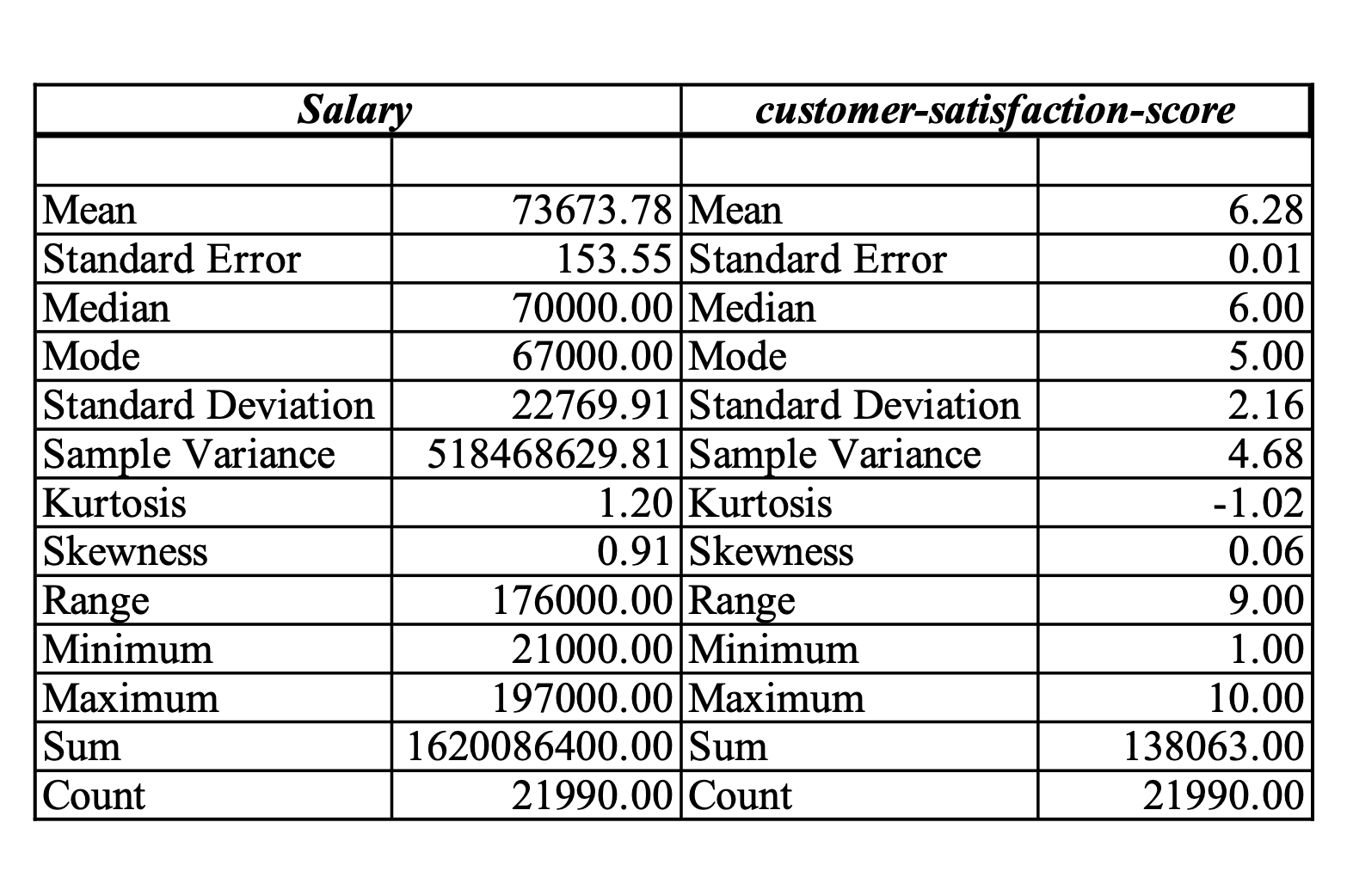
| **Variable** | **Description** |
| --- | --- |
| Sales\_Rep | It is a nominal variable, it shows the unique ID of sales representatives |
| Business | This is a categorical variable, describes about the two types of business the sales representatives working in such as Hardware, Software |
| Age | This is a quantitative variable of type discrete, it describes about the particular age group representatives involved in this data, we can observe that the people involved in this data having age in between 21 and 65 |
| Female | This is considered as a categorical variable as it describes whether the person is female or not, ‘1’ indicates as ‘true’ and ‘ 0’ indicates as ‘false’ |
| Years | It is a discrete/numerical variable, indicates the years of experience individual is having |
| College | It is a categorical variable, describes about the individual having college degree or not, indicates ‘Yes’ or ‘No’ |
| Personality | This variable comes under a categorical, describes the type of personality the individuals are having such as Analyst, Diplomat, Explorer, Sentinel |
| Certificates | This is a discrete/numerical data, describes about how many certificates individual having in the industry |
| Feedback | This is a continuous variable, describes the feedback of the individual |
| Salary | This is a discrete/numerical variable, indicates the salary of the sales representative, from observation the minimum salary is 21000 and maximum salary is 197000 |
| Customer-Satisfaction-Score | This is a discrete variable, describes about the rating given by the customer to an individual sales representative it ranges from 1-10 |

**3. Summary of Descriptive Statistics of Numerical Variables using EXCEL:**

A thorough overview of the most important statistical metrics and features of a dataset is provided by a summary of descriptive statistics using Excel. To get a general picture of the dataset, a summary of descriptive statistics on quantitative variables, both discrete and continuous, was performed.





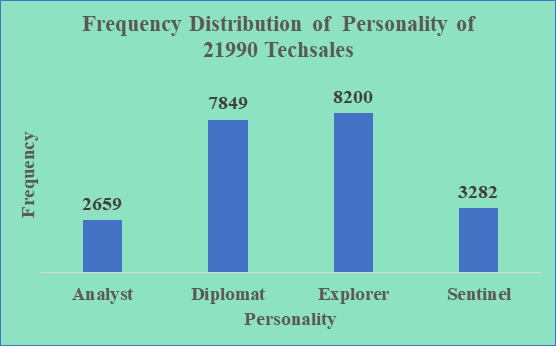


**4. Descriptive statistics of one or more categorical variable:**

Performing the descriptive statistics by considering one or more categorical variables by using excel we are calculating the frequency, percentage frequency and relative frequency of the variables.

| **4.1 Frequency Distribution of Personality Categorical Variables** |
| --- |

| **Personality** | **Frequency** | **% Frequency** | **Relative Frequency** |
| --- | --- | --- | --- |
| Analyst | 2659 | 12.09% | 0.121 |
| Diplomat | 7849 | 35.69% | 0.357 |
| Explorer | 8200 | 37.29% | 0.373 |
| Sentinel | 3282 | 14.92% | 0.149 |
| **Grand Total** | **21990** | **100.00%** | **1.000** |

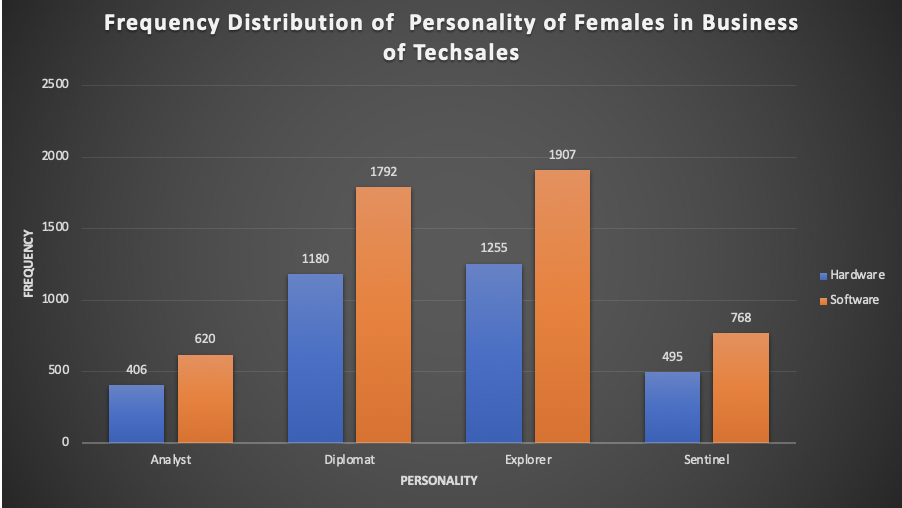
****

The distribution of personality types among sales reps is shown in the bar chart. Businesses can use this information to better identify the personality qualities that are most prevalent within its salesforce. If Explorer personalities, for instance, are more prevalent, it implies that the sales team may be made up of individuals who are daring, inquisitive, and eager to explore new chances. Companies can modify their management and training practices to take full advantage of these traits by recognizing the dominant personalities. For instance, diplomats, who are renowned for their diplomatic and compassionate disposition, can be particularly adept at developing lasting relationships with clients.Companies can allocate suitable sales reps to particular consumer segments by analyzing the personality mix, boosting the likelihood of successful sales outcomes.

| **4.2 Frequency Distribution of Personality, Business and Female Variables** |
| --- |

| **Frequency** | **Business** |  |  |
| --- | --- | --- | --- |
| **Personality** | **Hardware** | **Software** | **Grand Total** |
| Analyst | 406 | 620 | 1026 |
| Diplomat | 1180 | 1792 | 2972 |
| Explorer | 1255 | 1907 | 3162 |
| Sentinel | 495 | 768 | 1263 |
| **Grand Total** | **3336** | **5087** | **8423** |

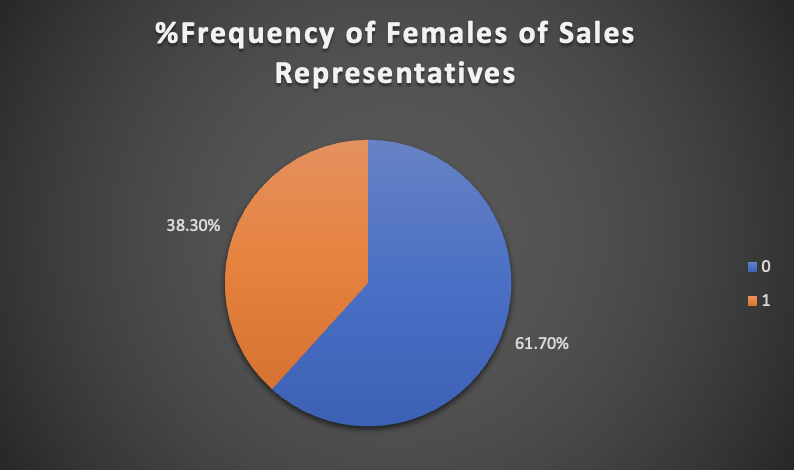
Businesses can detect possible alignment or mismatch between personality types and business requirements by understanding how personalities are distributed throughout various industries. For instance, because personality traits are more prevalent in the software sector than in hardware, there are chances for businesses to target and grow within the software market segment. As a result, businesses can access specialized skill sets, market knowledge, and customer insight unique to the software domain. This knowledge can help firms assign salespeople with the right personality types to different company areas for the best success.

****

|  |
| --- |

**4.3 % Frequency Distribution of Female of Sales Representatives**

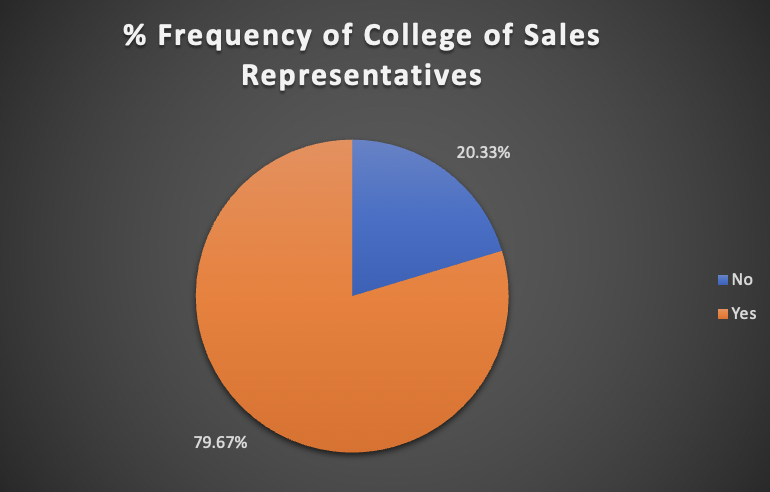
| **Female** | **% Frequency** |
| --- | --- |
| 0 | 61.70% |
| 1 | 38.30% |
| **Grand Total** | **100.00%** |

****

The representation of women among the sales people is shown in the pie chart. This information implies that the company is doing a better job of encouraging inclusion and gender diversity among its sales staff. In a sales force that is dominated by males, there may be a concentration of particular abilities or attributes if men are thought to perform well as sales representatives. This can suggest that certain traits, such aggressiveness or negotiation skill, are more common in men in the particular corporate setting. Utilizing these findings, companies may work to build a more diverse and inclusive sales team, which will improve performance, customer satisfaction, and brand perception.

**4.4 %Frequency Distribution of College of Sales Representatives**

| **College** | **% Frequency** |
| --- | --- |
| No | 20.33% |
| Yes | 79.67% |
| **Grand Total** | **100.00%** |

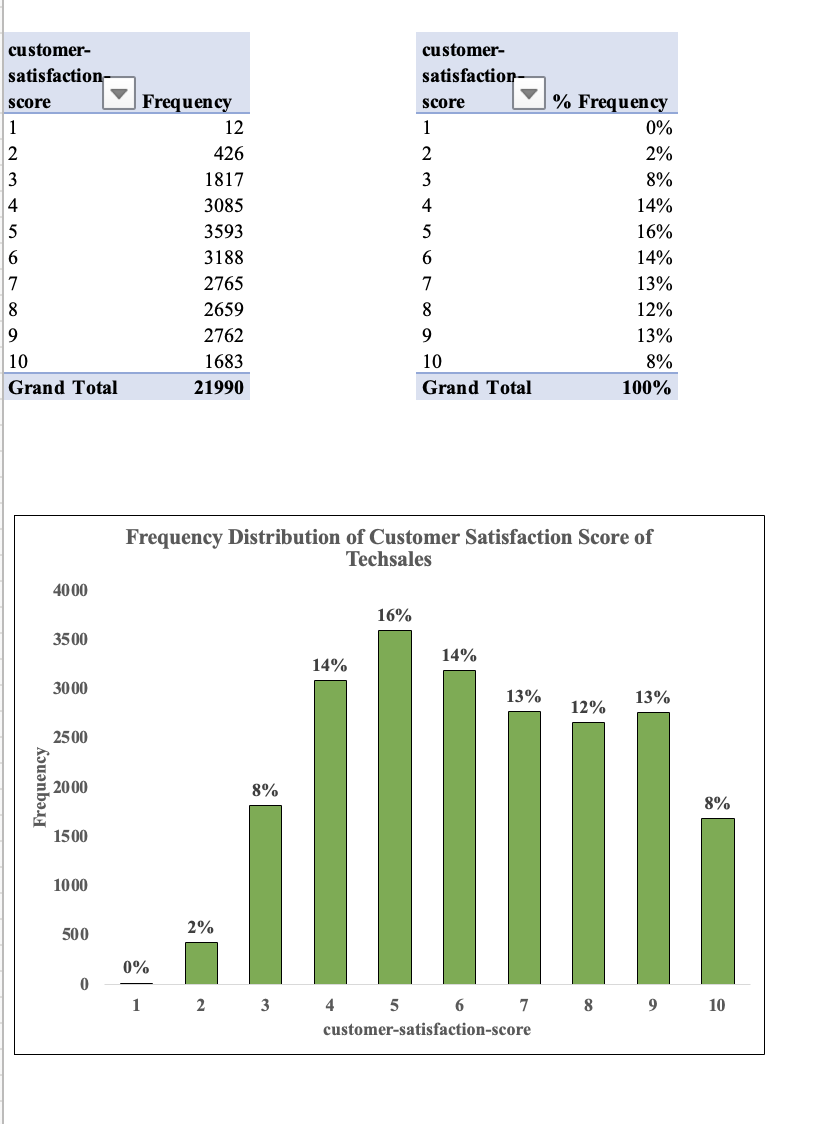
****

The pie chart shows that, of the sales representatives in the dataset, a sizable majority, or 79.67%, had a college degree. This suggests that the company employs a staff with more formal education, which provides a number of advantages such as a greater knowledge base, critical thinking abilities, and the capacity to adapt to challenging assignments. The increased proportion of salespeople with college degrees shows that they may have picked up a wide range of skills and knowledge during their educational experience. These qualities may include aptitudes for communication, problem-solving, research, and analysis. Businesses may use their sales team to the fullest extent possible and make use of their educational backgrounds to boost sales effectiveness by identifying and utilizing these skills.

**5. Descriptive statistics of one or more quantitative variables:**

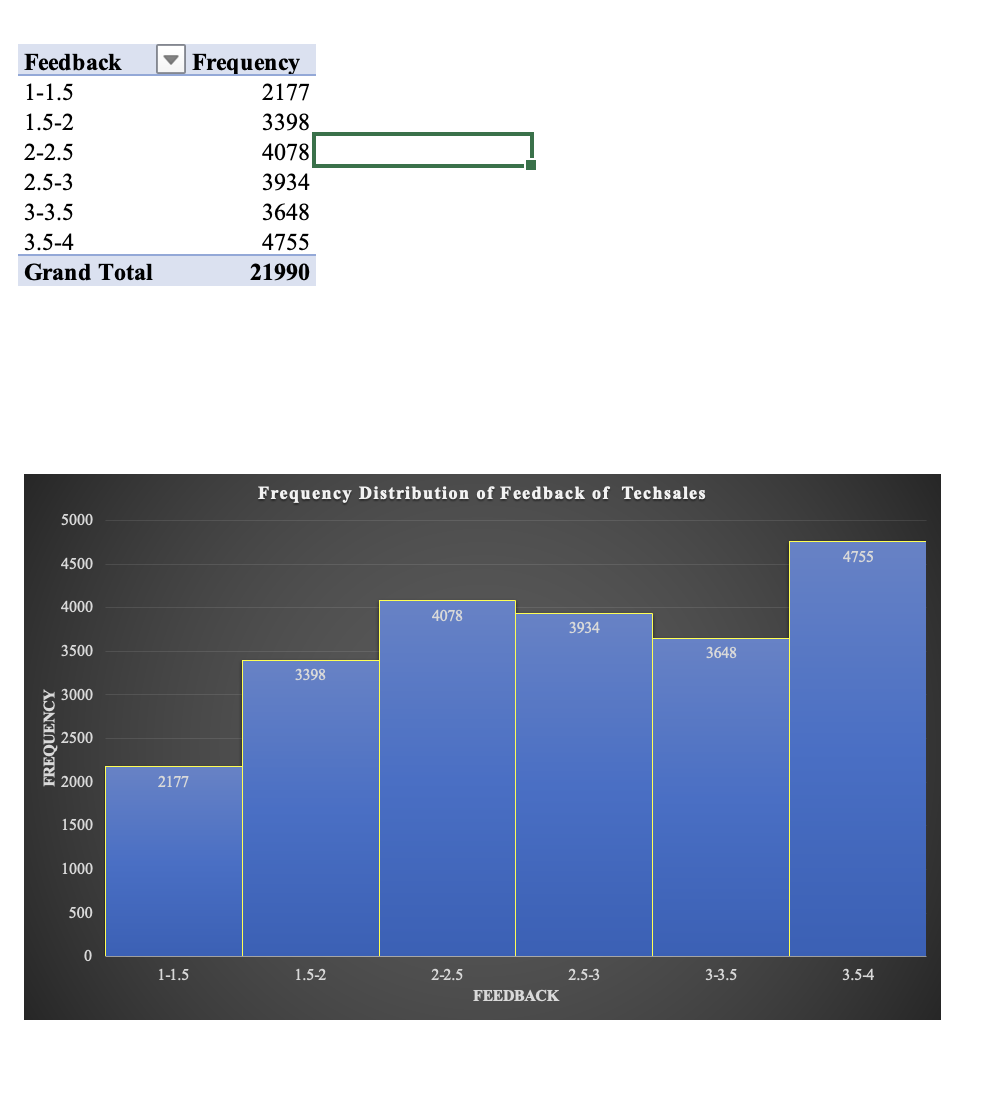
**5.1 Frequency Distribution of Customer Satisfaction Score of Discrete Variable and Age, Feedback of Continuous Variables**

| **customer-satisfaction-score** | **Frequency** | **% Frequency** |
| --- | --- | --- |
| 1 | 12 | 0% |
| 2 | 426 | 2% |
| 3 | 1817 | 8% |
| 4 | 3085 | 14% |
| 5 | 3593 | 16% |
| 6 | 3188 | 14% |
| 7 | 2765 | 13% |
| 8 | 2659 | 12% |
| 9 | 2762 | 13% |
| 10 | 1683 | 8% |
| **Grand Total** | **21990** | **100%** |

****

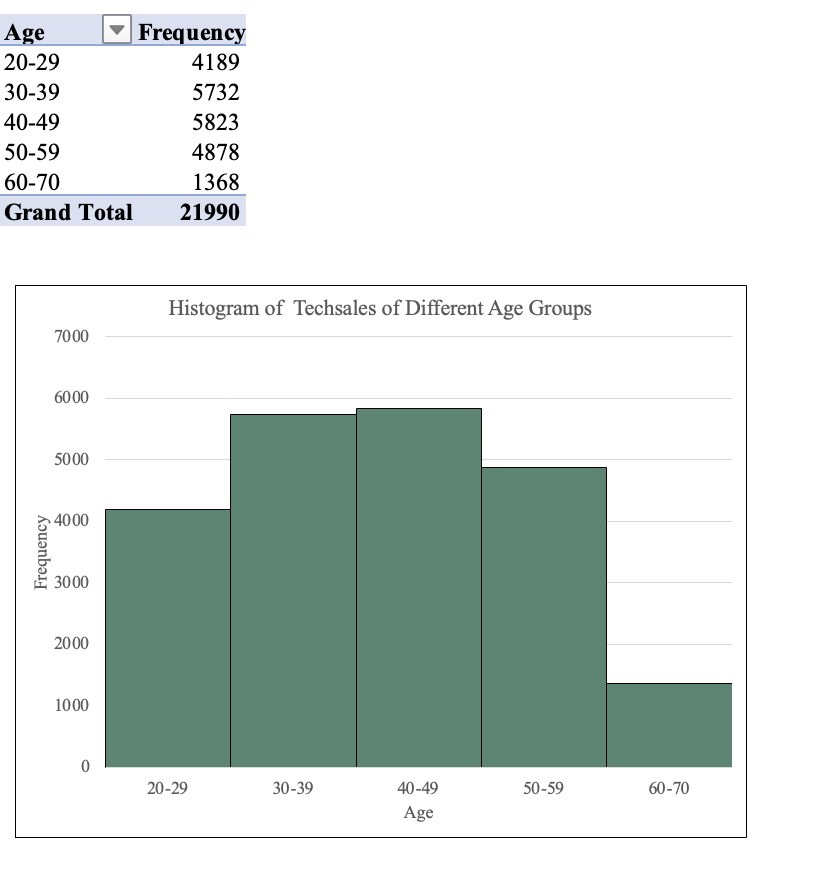
The above visualization shows the frequency distribution of customer satisfaction scores of TechSales. The bar graph is plotted between customer satisfaction scores ranging from 1 to 10 with the frequency of occurrences of each score. Most frequently rated score is 5 with the frequency of 3593 i.e, it is 16% of the total frequency and the lowest frequency being 12 i.e, 0% of the total with the customer score of 1. You can also observe the customer score with 3,10 having the same 8% of frequency. Setting objectives and monitoring advancement over time might be aided by using the distribution of customer satisfaction scores as a benchmark. Businesses can set reasonable goals and gauge their effectiveness in raising customer satisfaction by comparing current ratings to planned ones. This enables businesses to keep track of their progress, take appropriate action, and constantly improve their customer-centric strategies.

| **Feedback** | **Frequency** |
| --- | --- |
| 1-1.5 | 2177 |
| 1.5-2 | 3398 |
| 2-2.5 | 4078 |
| 2.5-3 | 3934 |
| 3-3.5 | 3648 |
| 3.5-4 | 4755 |
| **Grand Total** | **21990** |

****

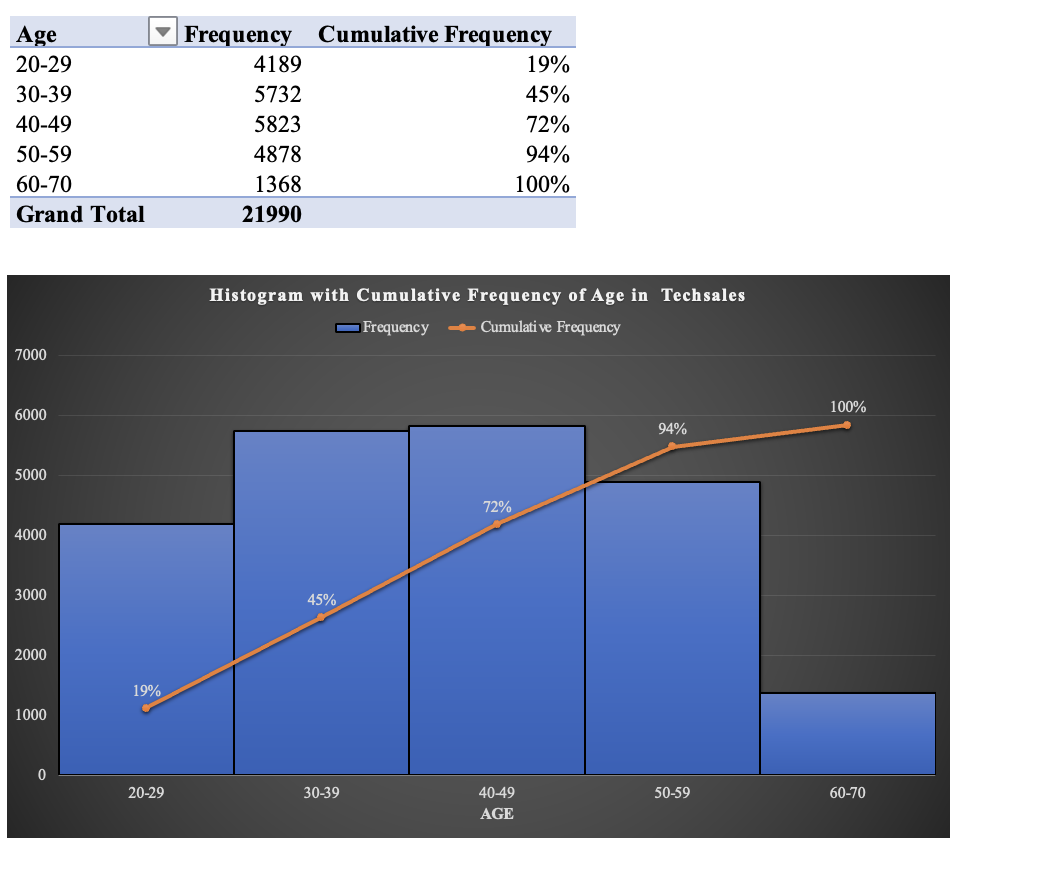
The values in the feedback variable are grouped together with the interval of 0.5. The above histogram represents the frequency distribution of a feedback for a TechSales company. Where we can see the feedback within the range of 3.5-4 has the highest frequency of 4755 and the lowest being the range of 1-1.5 with the frequency of 2177.Overall, this histogram offers valuable insights into the distribution of feedback scores received by the tech sales company highlighting the areas of improvement and focusing on the most common feedback range. According to the frequency distribution of the feedback scores, the majority of the scores are clustered at the upper end of the scale, which suggests that positive feedback is more common. This may indicate that consumers are typically happy with the company. There is still space for progress, though, by addressing the feedback from the lower ranges and using the positive feedback to fuel corporate expansion.

| **Age** | **Frequency** |
| --- | --- |
| 20-29 | 4189 |
| 30-39 | 5732 |
| 40-49 | 5823 |
| 50-59 | 4878 |
| 60-70 | 1368 |
| **Grand Total** | **21990** |

****

The above histogram represents the distribution of tech sales across various age groups. The highest frequency is observed in the age group of 40-49, indicating a significant number of tech sales from customers with the frequency of 5823. The age groups of 30-39 and 50-59 also exhibit relatively high frequencies, suggesting a substantial customer base in those demographics. Overall, this histogram provides a concise overview of the distribution of tech sales among different age groups, emphasizing the dominant age range of 40-49 for sales and highlighting potential opportunities for targeted marketing or growth in specific demographics.

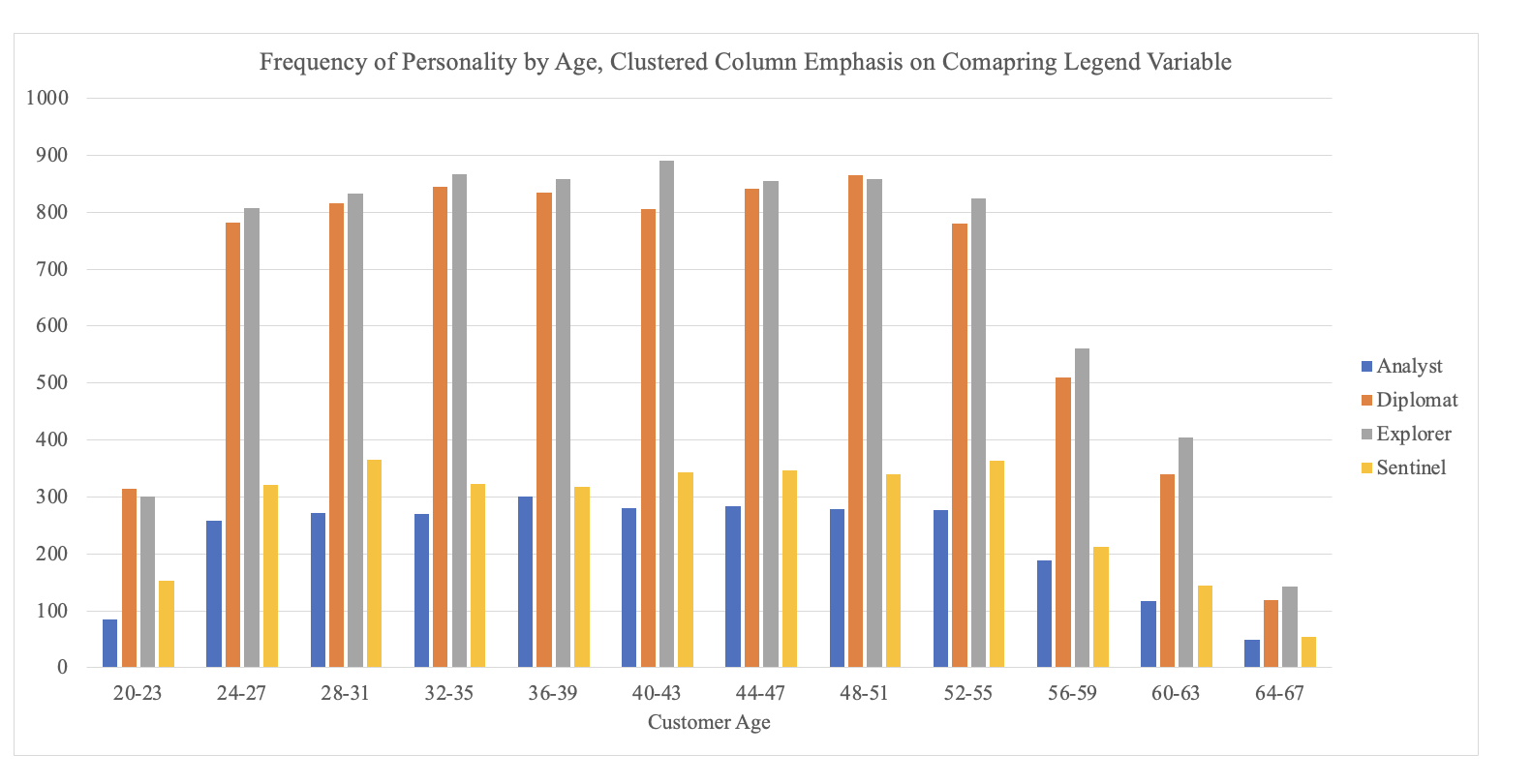
| **Age** | **Frequency** | **Cumulative Frequency** |
| --- | --- | --- |
| 20-29 | 4189 | 19% |
| 30-39 | 5732 | 45% |
| 40-49 | 5823 | 72% |
| 50-59 | 4878 | 94% |
| 60-70 | 1368 | 100% |
| **Grand Total** | **21990** |  |

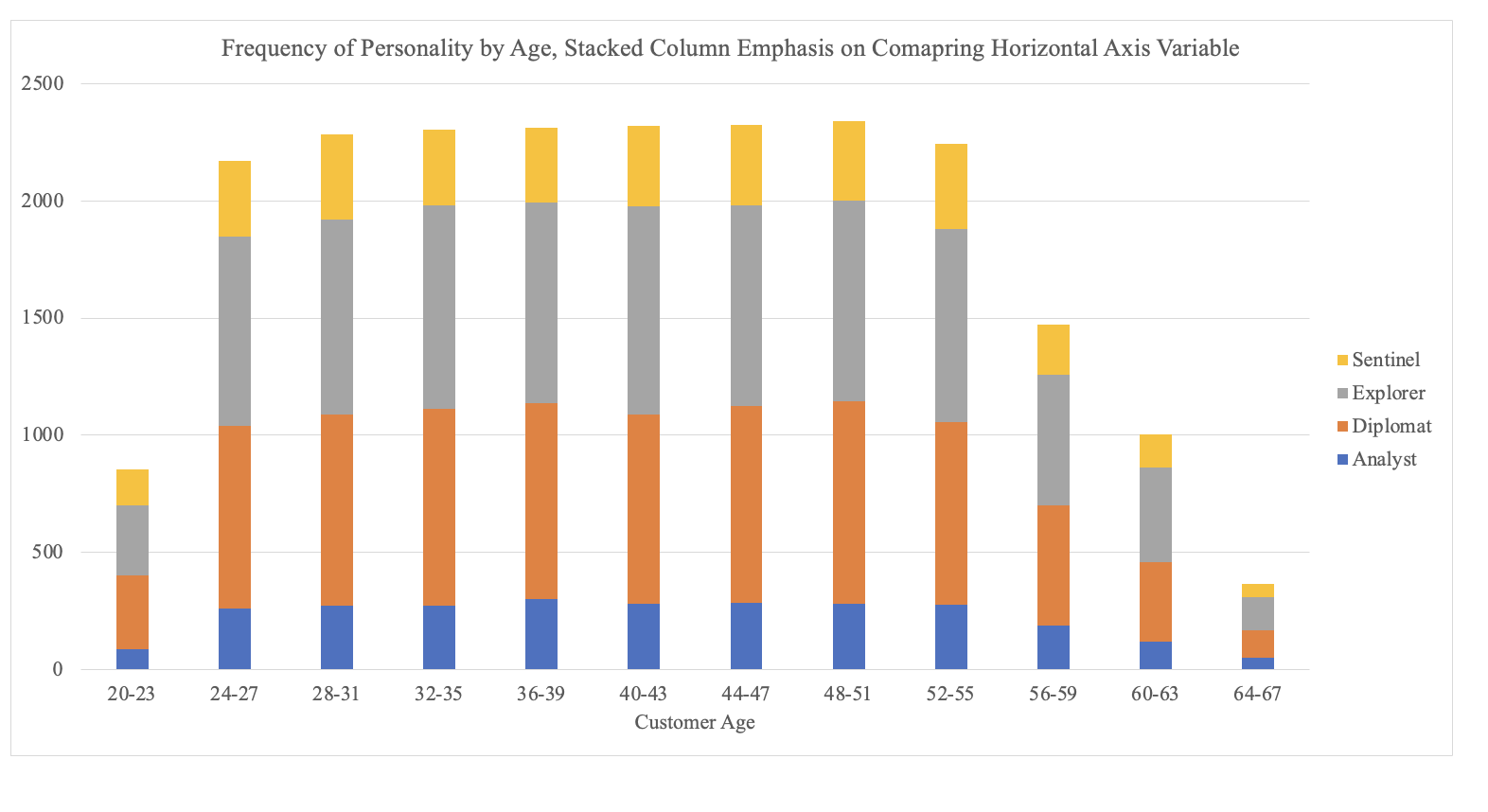
****

The above histogram illustrates the cumulative frequency distribution of age in the Tech Sales field, indicating the total number of individuals falling within various age groups. The age group of 40-49 exhibits the highest cumulative frequency of 72%. Overall, the histogram demonstrates a progressive rise in cumulative frequency from the youngest age groups to the high-aged category, reaching its peak at 60-70, followed by a gradual decline for the younger age groups. This data can be used by businesses to plan their workforce, find talent, and manage their operations. For instance, if the 40–49 age group predominates, it may be a sign that the organization needs to concentrate on finding and keeping talent in this age group, understanding their career ambitions, and offering opportunities for their professional progress.To show how frequencies gradually increase as you move through the age groupings, cumulative frequency for the age variable was provided. It is not essential or appropriate for all variables in every dataset, though. Depending on the features of the variables and the type of analysis being done, distinct variables may need various statistical measurements or presentation techniques.

**5.2 Frequency Distribution of Personality by Age**

| **Frequency** | **Personality** |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Customer Age** | **Analyst** | **Diplomat** | **Explorer** | **Sentinel** | **Grand Total** |
| 20-23 | 85 | 315 | 301 | 152 | 853 |
| 24-27 | 259 | 782 | 807 | 321 | 2169 |
| 28-31 | 271 | 815 | 833 | 365 | 2284 |
| 32-35 | 270 | 844 | 867 | 322 | 2303 |
| 36-39 | 300 | 835 | 859 | 318 | 2312 |
| 40-43 | 281 | 805 | 890 | 344 | 2320 |
| 44-47 | 283 | 841 | 855 | 346 | 2325 |
| 48-51 | 278 | 865 | 858 | 339 | 2340 |
| 52-55 | 277 | 780 | 824 | 364 | 2245 |
| 56-59 | 188 | 510 | 560 | 213 | 1471 |
| 60-63 | 117 | 339 | 404 | 144 | 1004 |
| 64-67 | 50 | 118 | 142 | 54 | 364 |
| **Grand Total** | **2659** | **7849** | **8200** | **3282** | **21990** |

****

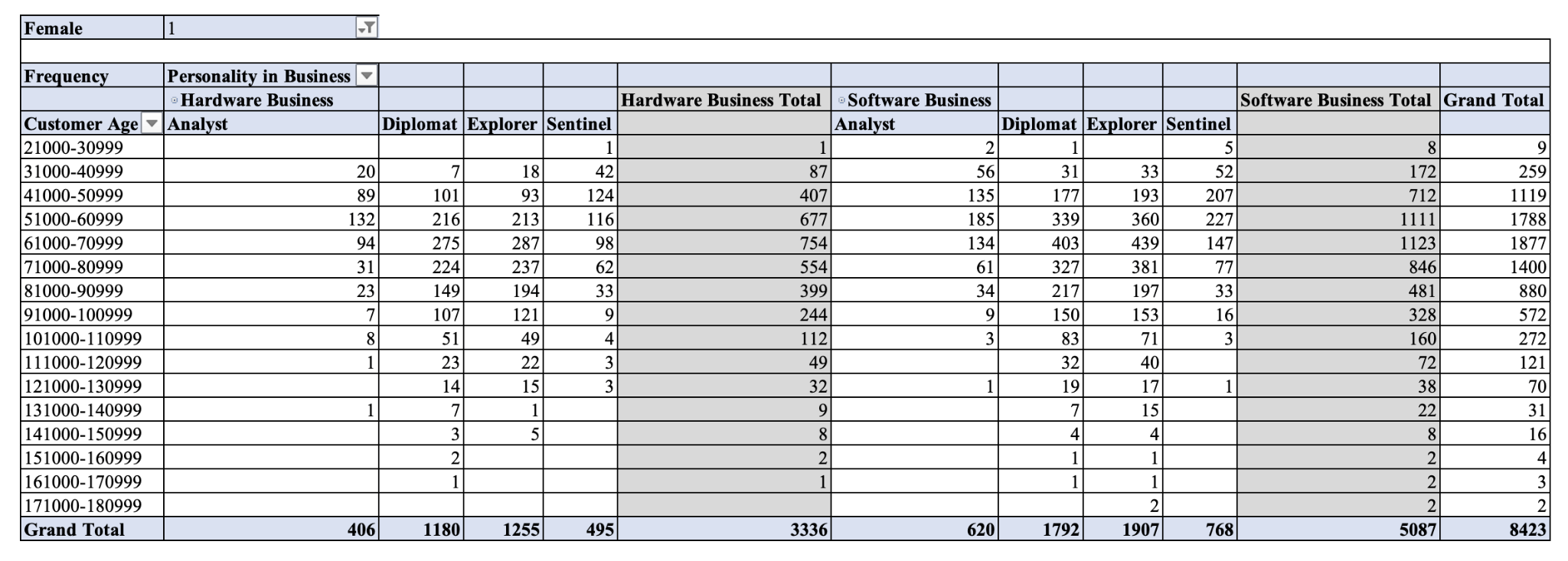
****

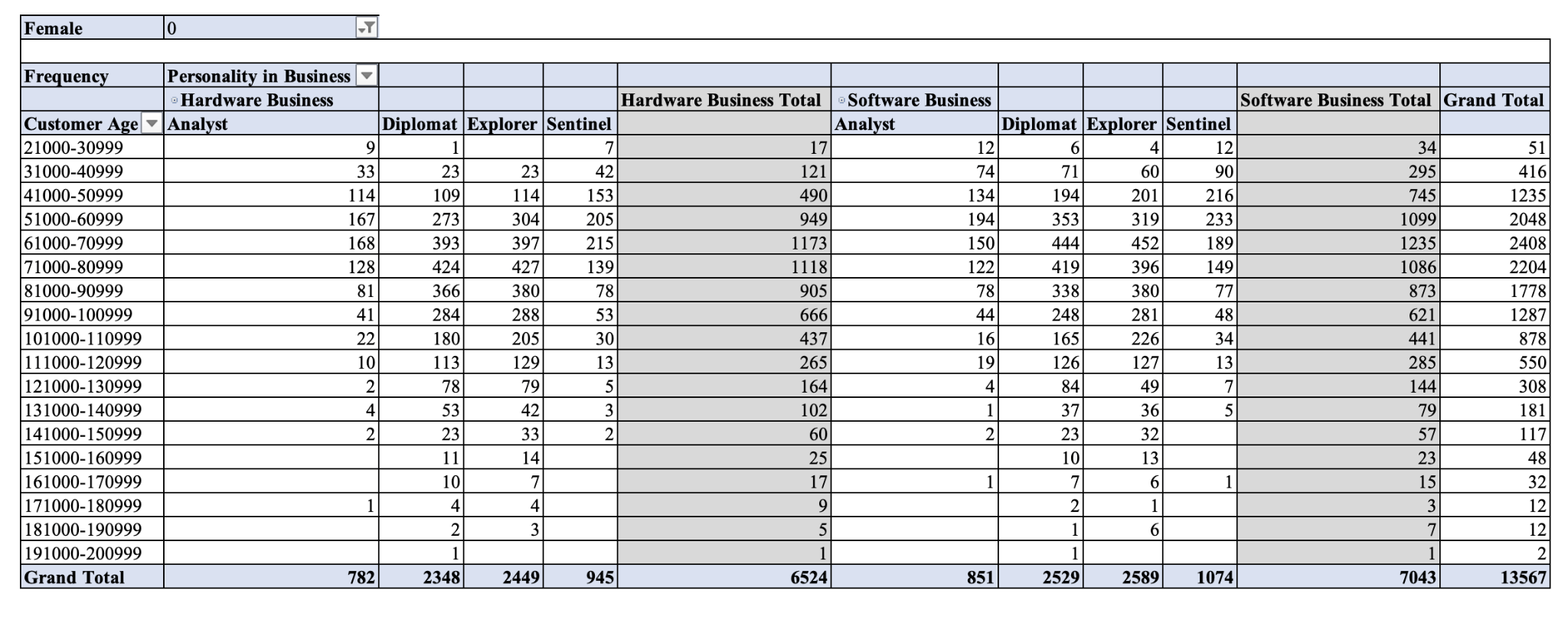
Both the above clustered and stacked charts convey similar meanings regarding the frequency distribution of personality types among customers of different age groups. However, the clustered chart provides better insights into the individual personality types of each age group compared to the stacked chart. In the clustered chart, the data is divided into separate clusters, allowing for a clearer view of the higher and lower frequencies within different age groups. Conversely, in the stacked chart, it is more challenging to compare the frequency distributions of the various age groups since the data is organized in a single stack. From both the graphs, the highest frequency for each personality type among age groups was 890 for Explorers in the 40-43 age group, while the lowest frequency was 50 for Analysts in the 64-67 age group, with overall decreasing trends observed for most personality types.By analyzing the frequencies and trends of personality types, businesses can develop targeted approaches to engage and cater to specific customer segments, ultimately improving customer satisfaction and driving business growth.

**6. Simpson's Paradox:**

A statistical phenomenon known as Simpson's Paradox happens when a pattern or connection shows up in various groups of data but reverses or vanishes when the groups are merged. It is a paradoxical finding that calls into question our preconceived notions about the facts. It develops when a hidden or confusing variable exists in the data. This variable affects both the independent and dependent variables, which might result in conclusions that are incorrect or inconsistent.

The below tables represent the simpson’s paradox phenomenon, in this dataset, there is a hidden variable associated with the "female" field. This variable has values of '0' and '1'. When the value is '0', it indicates that the sales representatives are male. This hidden variable influences other variables in the dataset, both independent and dependent variables, and should be taken into consideration when analyzing the data.

****

**** The tables show how various personality traits, separated by gender, affect sales effectiveness in the hardware and software industries. Businesses can determine which personality types—male or female—are more frequent in each business sector by evaluating the frequency distribution. In order to increase overall sales effectiveness, this study can be used to pinpoint efficient sales tactics, pinpoint the advantages of particular personality types, and optimize team composition. The business's ability to attract and engage female clients may be hampered by the excessive representation of male sales people. It's possible that different genders have different purchasing patterns, opinions, and requirements. A more diverse sales force, with balanced representation of men and women, offers a larger range of viewpoints, experiences, and ideas.

**7. Conclusion:**

As a result, the descriptive statistics analysis offers important details as well as insights regarding the dataset under examination. We now have a crystal-clear understanding of the conventional or average values in the dataset according to the summary statistics, which include calculations for the mean, median, and mode. These measurements serve as a point of comparison and aid in comprehending the data's primary tendency.Variability measurements like range, variance, and standard deviation have shed illumination on how evenly distributed the data is. The distribution patterns between the variables have been visually represented by the graphical representations, such as histograms, bar charts and line charts. Cross-tabulation has also been helpful in examining the connections between categorical variables and figuring out any connections or dependencies between them.