R Notebook

Code

Jenna Leali - Homework 2 - Demographic Dataset - Oct 3

Load Tidyverse and GGplot2

Hide

```
library(tidyverse)
library(ggplot2)
```

Read CSV File

Hide

```
dem_data <- read.csv("/Users/jenna/Desktop/demo.csv")
```

1. Review the data table and look at the data types, distributions, values, etc

View the first few rows of the dataset

Hide

```
head(dem_data)
```

Look at data types and structure of dataset

Hide

```
str(dem_data)
```

- Categorical variables like gender, education, marital_status, and occupation_group are stored as character vectors - numerical variables like age, credit_score, and home_build_year are appropriately stored as integers.

Get summary stats for numeric columns

Hide

```
summary(dem_data)
```

- Age: The age distribution appears to be somewhat right-skewed, as the mean age is slightly lower than the median age.
- Credit Score: Credit scores seem to be roughly normally distributed.
- Home Build Year: The majority of homes were built relatively recently, with a median build year of 1996. However, there are some very new homes in the dataset.

Look for missing values in dataset

Hide

```
colSums(is.na(dem_data))
```

- gender: 106 missing values.
- marital_status: 1,442 missing values.

Look at intial distributions

Hide

```
hist(dem_data$age, main = "Age Distribution", xlab = "Age")
hist(dem_data$credit_score, main = "Credit Score Distribution", xlab = "Credit Scor
e")
density_plot <- density(dem_data$home_build_year)
plot(density_plot, main = "Home Build Year Distribution", xlab = "Year")
```

Age Histogram

- Peak at 50-70: The peak in the histogram at the 50-70 age range suggests that there is a significant concentration of individuals in this age group.
- Distribution Shape: The histogram shape seems to be somewhat right-skewed, with fewer individuals in the younger age group (0-50).

Credit Score Histogram

- Credit Score Range: The credit scores range from 300 to 800, which is a common range for credit scores.
- Peaks in Frequency: There are peaks in the credit score distribution, particularly in the 400-450, 700-750, and 750-800 ranges. These peaks indicate that there are higher concentrations of individuals with

credit scores in these specific ranges.
- Balanced Distribution: The distribution appears to be relatively balanced, with a noticeable number of individuals across different credit score ranges.

Hide

```
table(dem_data$gender)
```

a. Detail any interesting patterns or issues you see.

- Missing Values: The age variable has 106 missing values, which might need to be handled. The marital_status variable has 1,442 missing values.
- Numeric Variables: The age variable ranges from 19 to 134, with a median age of 54. The credit_score variable ranges from 300 to 850, with a median score of 575. The home_build_year variable spans from 1970 to 2023, with a median year of 1996.
- Categorical Variables: categorical variables include gender, education, marital_status, occupation_group, address, city, and state.
- Outliers and Extreme Values: I will examine whether the extreme values in the age, credit_score, and home_build_year variables are valid or if they could be considered outliers.

2. Perform any necessary data cleaning and/or transformations. This could include, but not limited to, missing data, outliers, generation of new variables, binning, etc.

Missing Values: age and martial_status

Hide

```
dem_data$gender[is.na(dem_data$gender)] <- "Unknown"
dem_data$marital_status[is.na(dem_data$marital_status)] <- "Unknown"
```

- I decided to create the “Unknown” category in my dataset for the “gender” and “marital_status” columns because I encountered missing values in these variables.
- By introducing the “Unknown” category, I aimed to clearly represent cases where the data was incomplete, ensuring that these individuals are still included in my analyses while acknowledging the absence of specific gender or marital status information.

Outliers: numerical variables (age, creidt score, home build year)

Hide

```
boxplot(dem_data$age, main = "Age Distribution")
```

Hide

```
outliers_above_100 <- sum(dem_data$age > 100)
cat("Number of data points above 100 years old:", outliers_above_100, "\n")
```

Hide

```
dem_data <- dem_data[dem_data$age <= 100, ]
```

Hide

```
boxplot(dem_data$age, main = "Age Distribution")
```

- I first looked to see if the column 'age' has any outliers
- I found that there are outliers above the age 100. I then checked to see how many and found that there are 11 outliers.
- I decided to remove the outliers from the dataset because I believed them to be errors in the entry. I decided this as very few people live to such an age, and even fewer live to 120 years old. I decided that removing these individuals would increase the accuracy of my data.

Hide

```
boxplot(dem_data$credit_score, main = "Credit Score Distrubtion")
```

- There are no outliers for credit score distribution.

Hide

```
boxplot(dem_data$home_build_year, main = "Home Build Year Distribution")
```

- There are no outliers for home build year distribution.

Binning the age variable into categories: Young, Middle-aged, Elderly

Hide

```
dem_data$age_group <- cut(dem_data$age, breaks = c(0, 30, 60, Inf), labels = c("Youn
g", "Middle-aged", "Elderly"))
```

- I categorized the age category into three groups: “Young” for ages 0-30, “Middle-aged” for ages 31-60, “Elderly” for ages 61 and above.
- I decided to group people by age in my dataset to keep things simple and make it easier to see how different age groups might affect other variables.

Generation of new variables

Full Name

Hide

```
dem_data$full_name <- paste(dem_data$first_name, dem_data$last_name, sep = " ")
```

- I decided to create a new variable, “full_name,” by merging the “first_name” and “last_name” fields. This way, I now have a convenient variable that holds the complete names of the individuals.

Combine City, State

Hide

```
dem_data$location <- paste(dem_data$city, dem_data$state, sep = ", ")
```

- I combined the “city” and “state” variables in my dataset to create a new variable called “location.”. This way, I now have a consolidated location variable that I believe is better for geographical analysis.

Region Variable

Hide

```
dem_data$region <- case_when(
  dem_data$state %in% c("NY", "NJ", "PA", "CT", "MA", "VT", "NH", "ME", "RI") ~ "Nort
heast",
  dem_data$state %in% c("CA", "OR", "WA", "AK", "HI", "NV") ~ "West",
  dem_data$state %in% c("TX", "AZ", "NM", "OK") ~ "Southwest",
  dem_data$state %in% c("FL", "GA", "AL", "MS", "LA") ~ "Southeast",
  dem_data$state %in% c("IL", "MI", "WI", "OH", "IN", "KY") ~ "Midwest",
  TRUE ~ "Other"
)
```

- I decided to categorizing individuals by region. It allows for the assessment of regional credit score trends and disparities.

3. Summarize and/or aggregate the data table values in various ways with descriptive stats, counts, etc. over the entire dataset and by various groupings.

Hide

```
summary(dem_data$age)
summary(dem_data$credit_score)
```

Summary statistics

- The age distribution appears relatively broad, with individuals ranging from 19 to 97 years old. The median age of 54 suggests that most people in the dataset fall around this age.
- For credit scores, the distribution is pretty balanced, ranging from a minimum of 300 to a maximum of 850. The median and mean credit scores are both 575, indicating that credit scores are evenly spread across this range.
- The quartile values provide a sense of the data's spread, showing that the majority of individuals have credit scores between 441 and 710.

Count of categorical variables

Hide

```
table(dem_data$marital_status)
```

- From the marital status distribution, I found that the majority of individuals are categorized as Married. There's also a smaller group classified as Single and a notable portion falls under Unknown marital status.

Summary Stats by grouping

Hide

```
dem_data %>%
  group_by(education) %>%
  summarise(
    Mean_Credit_Score = mean(credit_score),
    SD_Credit_Score = sd(credit_score)
  )
```

- From the data, I can see that credit scores differ across various education levels. Individuals with a “Grad Degree” tend to have the highest average credit score (580.41), while those with a “Bach Degree” follow closely (572.82).
- The standard deviations in credit scores are relatively consistent across the different education categories, indicating moderate variability within each group.
- “Less than HS Diploma” holders exhibit the highest variability, but their mean credit score (579.54) is notably higher than the overall dataset’s mean.
- These findings suggest that education might play a role in creditworthiness, with higher educational attainment associated with slightly higher average credit scores.

Aggregating data by multiple variables: calculate the average age for different combinations of “marital_status” and “gender”

Hide

```
aggregate(dem_data$age, by = list(dem_data$marital_status, dem_data$gender), FUN = me
an)
```

- By using the aggregate function, I uncovered some insights about the mean age within various marital status and gender combinations.
- It seems that individuals with an “Unknown” marital status tend to have notably higher mean ages, particularly “Unknown” females with an average age of approximately 64.51 years.
- “Married” males also exhibit a higher mean age at around 53.84 years, suggesting that marriage might be associated with slightly older age.
- On the other hand, “Single” females and males tend to have lower mean ages, hovering around 40.29 years.

Hide

```
dem_data %>%
  group_by(education, marital_status) %>%
  summarise(
    Mean_Credit_Score = mean(credit_score)
  )
```

- Insights into how credit scores vary across different combinations of education levels and marital statuses:
- “Single” individuals tend to exhibit higher mean credit scores within each education level. For example, individuals with “Bach Degree” education and a “Single” marital status have an average credit score of around 596.94, suggesting that being single might be associated with slightly better creditworthiness within this educational category.
- Individuals with “Less than HS Diploma” and a “Single” status also have notably higher credit scores,

with a mean of approximately 594.41, compared to their “Married” counterparts. Maybe these individuals have higher credit scores because they did not need to take loans for college or they do not share a credit socre with their spouse.
- These findings suggest that both education and marital status are important factors influencing credit scores.

Hide

```
dem_data %>%
  group_by(gender, occupation_group) %>%
  summarise(
    Count = n()
  )
```

- From this data, I’ve gained insights into the distribution of individuals across different gender and occupation groups.
- For “Female” individuals, it’s interesting to note that the largest groups are in “Other” and “Professional” occupations, with 635 and 646 individuals.
- On the other hand, for “Male” individuals, the most significant presence is in “Blue Collar” and “Management” categories, with 2,310 and 1,569 individuals.
- The “Retired” occupation group also appears substantial for both genders.
- It’s worth considering the “Unknown” gender category, where the numbers are notably smaller across various occupation groups.

Hide

```
dem_data %>%
  group_by(gender, occupation_group) %>%
  summarise(
    Count = n(),
    Mean_Credit_Score = mean(credit_score)
  )
```

- Upon examining the data, I’ve identified some patterns regarding credit scores within different gender and occupation groups.
- It appears that Male individuals in Management occupations tend to have higher average credit scores, which could indicate that their occupation is associated with better creditworthiness.
- On the other hand, Male individuals in Blue Collar occupations, while being a substantial group, exhibit lower average credit scores, suggesting that this occupational category might be associated with slightly lower creditworthiness.

Hide

```
dem_data %>%
  group_by(home_build_year) %>%
  summarise(
    Q1_Credit_Score = quantile(credit_score, 0.25),
    Median_Credit_Score = quantile(credit_score, 0.5),
    Q3_Credit_Score = quantile(credit_score, 0.75)
  )
```

- There is noticeable variation in credit scores across different home build years, with some years having a wider spread, indicating greater variability, while others have a narrower range, suggesting a more consistent distribution.
- Historical trends can be observed in credit scores, with years in the late 1970s to early 1980s, like 1980, tending to have higher median credit scores. This may be influenced by economic factors or lending practices during that period. Years in the early 1990s, such as 1995 and 1996, show a dip in median credit scores.
- Some years, such as 1977, 1987, and 2012, exhibit notably high Q3 credit scores, showing the presence of individuals with excellent credit scores. This could be linked to specific economic conditions or demographic shifts in those years.
- Years like 2003 and 2016 display relatively stable Q1, median, and Q3 credit scores, suggesting consistent creditworthiness for individuals whose homes were built during those years.
- Years with a wide range between Q1 and Q3, like 2008 and 2015, could be associated with greater variability, possible implying instability in credit scores.

Cross tabulations: cross-tabulate “education” and “marital_status

Hide

```
table(dem_data$education, dem_data$marital_status)
```

- I observed that individuals with a Bach Degree are predominantly married, with 1,631 individuals.
- Those with a Grad Degree exhibit a similar pattern, with the majority being married 1,036 individuals.
- Individuals with a HS Diploma are more diverse, with a substantial number of married 2,185 and single 274 individuals, as well as a significant number of individuals with an unknown marital status 568.
- People with Less than HS Diploma have a more evenly spread distribution across marital statuses, with 487 married, 79 single, and 162 individuals of unknown marital status.
- Those with Some College education have a higher number of married 2,828 and single 407 individuals, with a smaller group of unknown marital status 373.

4. Leveraging the analyses in steps 1-3, create at least four different plots over

variables you finding interesting to include univariate and multivariate (covariation) analyses. Make sure the plots are customized appropriately with labels, titles, colors, and themes.

Barplot of credit score by education level

Hide

```
ggplot(dem_data, aes(x = education, y = credit_score, fill = education)) +
  geom_boxplot() +
  labs(title = "Credit Score by Education Level",
       x = "Education Level",
       y = "Credit Score") +
  theme_minimal() +
  theme(legend.position = "none")
```

- The choice of a boxplot is suitable for visualizing the distribution of a continuous variable (credit score) within different categorical groups (education levels). It displays key summary statistics such as the median, quartiles, and potential outliers, making it easy to compare distributions.
- Across all education levels, there is a consistent median credit score around 560
- However, the wide credit score ranges indicate significant variability in credit scores within each education category.
- These patterns suggest that education, while a contributing factor, does not entirely account for the variation in credit scores. Other demographic factors, such as age, income, or gender, likely play a role in influencing credit scores within the provided dataset.

Hide

```
sampled_data <- dem_data[sample(nrow(dem_data), 1000), ]
ggplot(sampled_data, aes(x = age, y = credit_score, color = gender)) +
  geom_point() +
  labs(title = "Age vs. Credit Score (Sampled Data)",
       x = "Age",
       y = "Credit Score") +
  theme_minimal()
```

- The choice of a scatterplot is suitable for assessing the correlation or patterns between two continuous variables (age and credit score) and how gender affects this relationship. This plot allows for the

visualization of individual data points.
- The preponderance of green dots (males) suggests a larger representation of males in the dataset, while red dots (females) are less prominent.
- The concentration of data points within the 50-70 age range suggests that a substantial portion of the sampled data corresponds to individuals in this age group.
- The scatterplot emphasizes the wide distribution of data points, indicating that credit scores vary across age groups. This may suggest that age alone is not the sole determinant of credit scores in the dataset.

Hide

```
ggplot(dem_data, aes(x = gender, y = credit_score, fill = gender)) +
  geom_boxplot() +
  labs(title = "Credit Score by Gender",
       x = "Gender",
       y = "Credit Score") +
  theme_minimal() +
  theme(legend.position = "none")
```

- I opted for a boxplot to visually present how credit scores are distributed by gender within the dataset. Boxplots are an effective choice for comparing the distribution of a continuous variable (in this case, credit score) across various gender categories.
- When examining the boxplots, it is evident that the median credit scores for each gender category are relatively similar. This suggests that, on average, there is no substantial gender-based difference in credit scores within the dataset.
- The variability within each gender category is represented by the height and spread of the boxes. While the credit score range for the “unknown” gender category is wider, indicating greater variability, there is still substantial overlap with the ranges of the other two genders.

Hide

```
ggplot(dem_data, aes(x = region, y = credit_score, fill = marital_status)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Credit Score by Region and Marital Status",
       x = "Region",
       y = "Credit Score") +
  theme_minimal()
```

- I used a grouped bar chart to visually represent credit scores based on regions and marital status. In this plot, regions are on the x-axis, credit scores are on the y-axis, and the bars are color-coded by marital status. A grouped bar chart is a suitable choice for comparing the distribution of a continuous variable (credit score) across multiple categories (regions) while considering another categorical variable (marital status).
- In the “Midwest,” credit scores varied by marital status, with “Married” individuals boasting the highest scores at 850, followed by “Single” individuals at 810, and “Unknown” individuals with the lowest

scores at 780.
- In contrast, the “Northeast” had only one bar for “Married” individuals, reflecting a credit score of 770.
- The “Other” region exhibited variations, with “Married” individuals having a credit score of 800, “Unknown” individuals at 810, and no representation for “Single” individuals.
- The “Southeast” region was unique, as all marital statuses displayed the same high credit score of 850.
- Overall, the grouped bar chart reaffirms the prevalence of high credit scores within the dataset while revealing interesting variations across different regions and marital statuses.

Hide

```
ggplot(dem_data, aes(x = age_group, y = credit_score, fill = education)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Credit Score by Age Group and Education Level",
       x = "Age Group",
       y = "Credit Score") +
  theme_minimal()
```

- I noticed that all of the bars representing credit scores are consistently above 800, and there is a relatively even distribution across different age groups and education levels. This observation suggests that, on average, individuals in my dataset tend to maintain high credit scores, regardless of their age or educational background.
- I noticed that the “Middle-aged” category exhibits a more uniform distribution of credit scores compared to the “Young” and “Elderly” groups. This may indicate that individuals in the “Middle-aged” group tend to maintain relatively consistent credit scores across different education levels.

Hide

```
ggplot(dem_data, aes(x = marital_status, y = credit_score, fill = occupation_group))
+
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Credit Score by Marital Status and Occupation Group",
       x = "Marital Status",
       y = "Credit Score") +
  theme_minimal()
```

- The “Farmer” occupation group consistently shows the lowest credit scores across all marital status categories—married, single, and unknown.
- Within the “Married” category, all occupation groups, excluding “Farmer,” cluster around a relatively high credit score of around 820. This suggests that marriage generally correlates with higher credit scores, regardless of the occupation.
- In the “Single” category, there is slightly more variation, with the “Management” and “Retired” occupation groups exhibiting slightly lower credit scores compared to the others.
- In the “Unknown” gender category, credit scores follow a pattern similar to the “Married” category but with a touch more variation.
- These insights show the role of both marital status and occupation in influencing credit scores, with

"Farmer" being an outlier with consistently lower scores.

5. Summarize your interpretation of the overall results of your demographic analysis, discussing any interesting insights or trends you discovered. Posit what could be done with your analysis results—could this demographic data lead to actionable insights?

- The data highlights a consistent trend where individuals with higher educational attainment tend to have slightly higher mean credit scores. This insight can lead to the development of tailored financial education programs, helping those with lower education levels improve their credit management skills.
- While the data indicates that "Single" individuals tend to exhibit higher mean credit scores within each education level, there's a contrast when considering occupation groups. The lower credit scores among "Single" individuals in the occupation-specific analysis might be due to the influence of occupation on creditworthiness. This suggests that occupation and marital status collectively impact credit scores and could serve as a basis for lenders to fine-tune their risk assessment models, providing more nuanced loan offerings based on both variables.
- The data showcases variations in credit scores across different regions and marital statuses. This insight can inform economic policy decisions, potentially leading to targeted economic stimulus measures in regions with consistently lower credit scores.
- The observation that credit scores are relatively even across different age groups and education levels suggests that, on average, individuals maintain high credit scores regardless of these demographics. It may imply that traditional factors such as age and education level might not be the primary drivers of credit scores in the dataset.
- The data reveals variations in credit scores across different gender and occupation groups. This finding could lead to insights for human resources and inclusion initiatives.
- In conclusion, the dataset contains a diverse group of individuals with various credit scores, impacted by factors such as education, age, gender, occupation, region, and marital status. It appears that all of the variables play a role in credit scores and it truly depends on all of the variables put together.