Session 5

Victor Benito Garcia Rocha

2024-07-07

Part I. Load the libraries and import the excel file

1.1 Rename the dataset to CCC, keep only the following variables: Annual_income, Type_income, EDUCATION, Marital_status. Print the header of CCC.

```r
library(readxl)
library(tidyverse)
library(ggplot2)
Credit_card_costumers <- read_excel(
  "Credit_card_costumersV2.xlsx"
)

Credit_card_costumers
```

```
## # A tibble: 1,548 x 19
##     Ind_ID GENDER Car_Owner Propert_Owner CHILDREN Annual_income Type_Income
##      <dbl> <chr>  <chr>     <chr>            <dbl> <chr>         <chr>
##  1 5008827 M      Y         Y                    0 180000 USD    Pensioner
##  2 5009744 F      Y         N                    0 315000 USD    Commercial ass~
##  3 5009746 F      Y         N                    0 315000 USD    Commercial ass~
##  4 5009749 F      Y         N                    0 <NA>          Commercial ass~
##  5 5009752 F      Y         N                    0 315000 USD    Commercial ass~
##  6 5009753 <NA>   Y         N                    0 315000 USD    Pensioner
##  7 5009754 F      Y         N                    0 315000 USD    Commercial ass~
##  8 5009894 F      N         N                    0 180000 USD    Pensioner
##  9 5010864 M      Y         Y                    1 450000 USD    Commercial ass~
## 10 5010868 M      Y         Y                    1 450000 USD    Pensioner
## # i 1,538 more rows
## # i 12 more variables: EDUCATION <chr>, Marital_status <chr>,
## #   Housing_type <chr>, Birthday_count <dbl>, Employed_days <dbl>,
## #   Mobile_phone <dbl>, Work_Phone <dbl>, Phone <dbl>, EMAIL_ID <dbl>,
## #   Type_Occupation <chr>, Family_Members <dbl>, `Debit card` <chr>
```

```r
getwd()
```

```
## [1] "C:/Users/Administrador/Desktop/BusinessAnalyticsITESM"
```

```r
CCC <- Credit_card_costumers %>%
  select(Annual_income, Type_Income, EDUCATION, Marital_status)
head(CCC)
```

```
## # A tibble: 6 x 4
##   Annual_income Type_Income          EDUCATION        Marital_status
##   <chr>         <chr>                <chr>            <chr>
## 1 180000 USD    Pensioner            Higher education Married
## 2 315000 USD    Commercial associate Higher education Married
## 3 315000 USD    Commercial associate Higher education Married
## 4 <NA>          Commercial associate Higher education Married
## 5 315000 USD    Commercial associate Higher education Married
## 6 315000 USD    Pensioner            Higher education Married
```

Part II. Exploring the dataset

2.1 Get a summary of all of the variables for CCC dataset to identify their characteristics. Verify that the variable "Annual_income" is a numeric one.

```
summary(CCC)
```

```
##  Annual_income       Type_Income         EDUCATION         Marital_status
##  Length:1548        Length:1548        Length:1548        Length:1548
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
class(CCC$Annual_income)
```

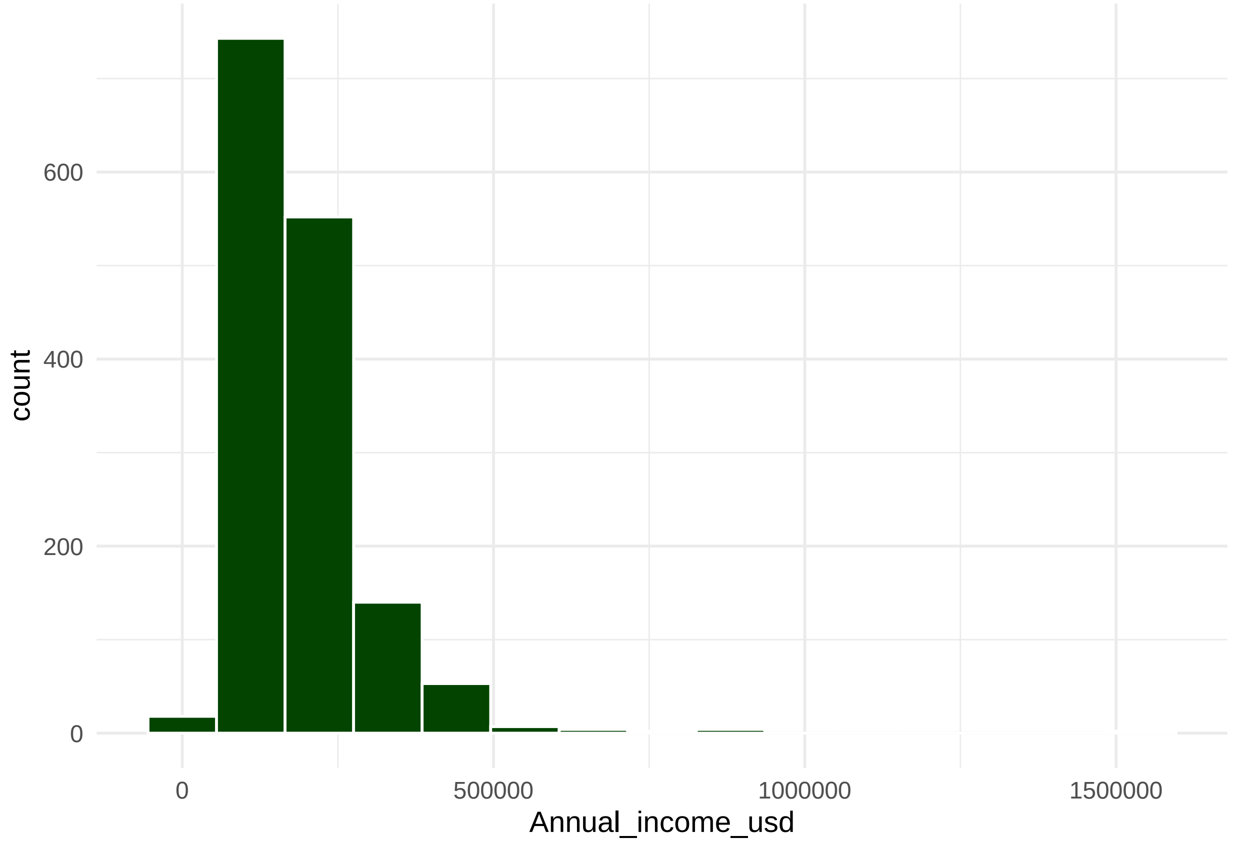
```
## [1] "character"
```

2.2 "Annual_income" does not seem to be numeric. Since you want to present a statistical summary of it as well as a graph to describe it, you need to change it to a numeric variable. First, you need to use the str_remove_all to delete the "USD" word (create a variable for this: Annual_income_trimmed), then, make it numeric (creare another variable: Annual_income_usd). After you solve that, get a histogram for that variable with labels in each axis and main title.

```
CCC <- CCC %>%
  mutate(Annual_income_trimmed = (str_remove(Annual_income, "USD"))) %>%
  mutate(Annual_income_usd = as.numeric(Annual_income_trimmed)) %>%
  select(Annual_income_usd, Type_Income, EDUCATION, Marital_status)

CCC %>%
  ggplot(aes(x = Annual_income_usd)) +
  geom_histogram(color = "white", fill = "#034400", bins = 15) +
  theme_minimal()
```

```
## Warning: Removed 23 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



What can you say about the skewness in this graph? Explain

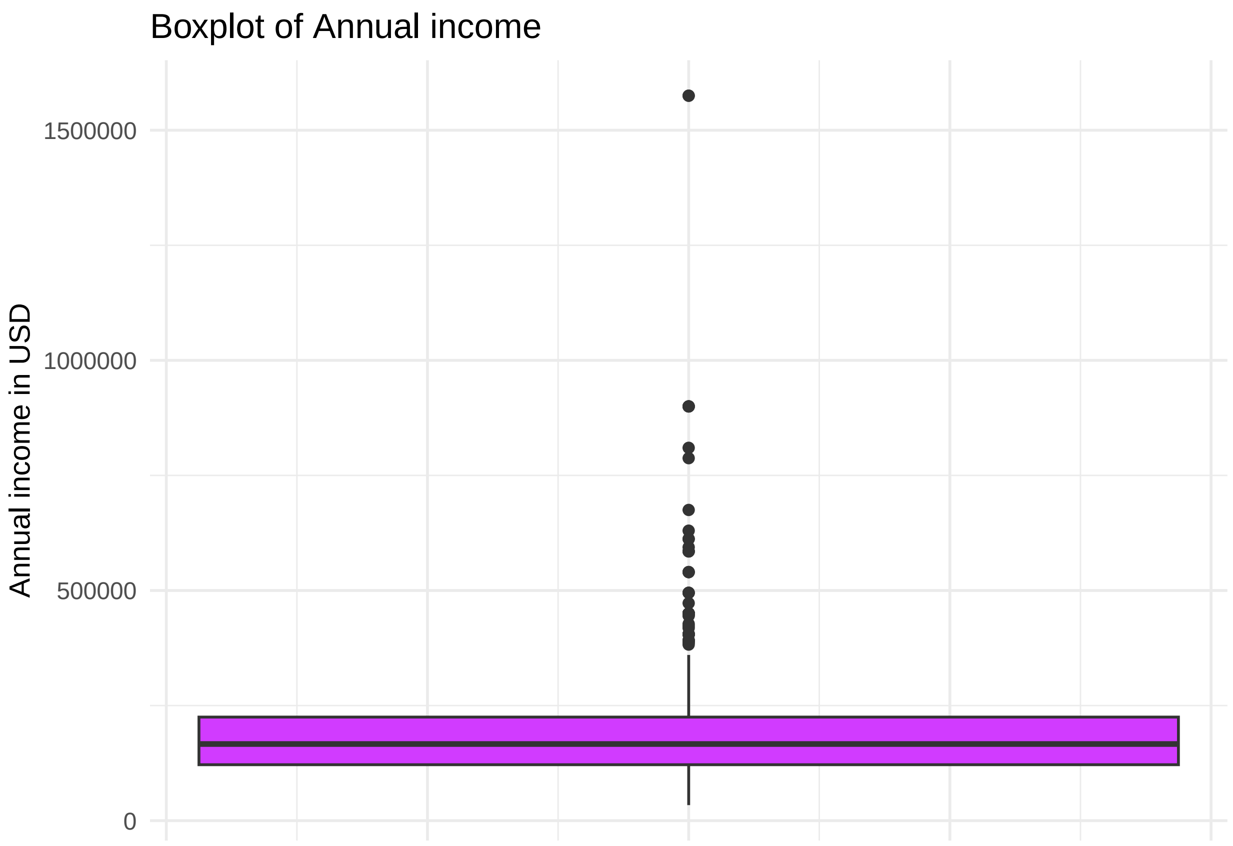
The majority of the data is concentrated on the left side of the histogram. This indicates that most of the people have lower annual incomes, while a few individuals have **much higher** incomes.

Part III. Detecting and dealing with outliers

3.1 Build a boxplot for the Annual_income_usd variable to identify potential outliers.

```
CCC %>%
  ggplot(aes(y = Annual_income_usd)) +
  geom_boxplot(fill = "#d13bff") +
  theme_minimal() +
  theme(axis.text.x = element_blank()) +
  labs(y = "Annual income in USD", title = "Boxplot of Annual income")
```

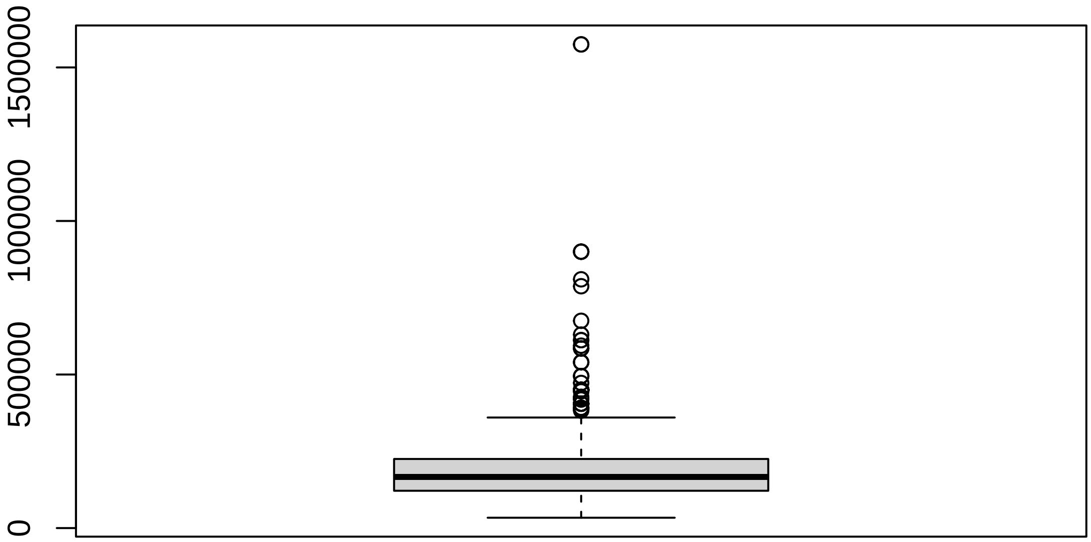
```
## Warning: Removed 23 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



Do you notice any outliers in the dataset? Answer here

The boxplot clearly shows the presence of several outliers in the data, with a significant number of individuals earning much higher incomes than the majority, contributing to the last right-skewed distribution.

```
# Identify the outliers (there are 73 in total):
boxplot(CCC$Annual_income_usd)$out
```



```
##  [1]  450000  450000  450000  472500  540000  540000  450000  391500  391500
## [10]  391500  675000  585000  585000  450000  450000  450000  450000  450000
## [19]  450000  445500 1575000 1575000  900000  450000  450000  423000  450000
## [28]  540000  450000  495000  612000  427500  612000  450000  787500  450000
## [37]  594000  585000  495000  387000  450000  900000  382500  450000  900000
## [46]  405000  405000  445500  450000  450000  450000  450000  405000  900000
## [55]  630000  450000  418500  450000  405000  405000  405000  495000  450000
## [64]  387000  810000  391500  405000  450000  405000  450000  450000  405000
## [73]  450000
```

```
# Create an object ("out") that will keep the observations considered outliers. After that, create "out_
out <- boxplot.stats(CCC$Annual_income_usd)$out
out_index <- which(CCC$Annual_income_usd %in% c(out))
CCC[out_index, ]
```

```
## # A tibble: 73 x 4
##    Annual_income_usd Type_Income          EDUCATION                Marital_status
##                <dbl> <chr>                <chr>                    <chr>
##  1            450000 Commercial associate Secondary / secondary ~ Married
##  2            450000 Pensioner            Secondary / secondary ~ Married
##  3            450000 Commercial associate Secondary / secondary ~ Single / not ~
##  4            472500 Pensioner            Higher education         Married
##  5            540000 Commercial associate Higher education         Married
##  6            540000 Commercial associate Higher education         Married
##  7            450000 Commercial associate Higher education         Separated
##  8            391500 Working              Secondary / secondary ~ Single / not ~
##  9            391500 Working              Secondary / secondary ~ Single / not ~
## 10            391500 Working              Secondary / secondary ~ Single / not ~
## # i 63 more rows
```

3.2 Find and show any observation with NA's in Annual_income_usd.

```
CCC[is.na(CCC$Annual_income_usd), ]
```

```
## # A tibble: 23 x 4
##    Annual_income_usd Type_Income          EDUCATION               Marital_status
##                <dbl> <chr>                <chr>                   <chr>
##  1                NA Commercial associate Higher education        Married
##  2                NA Working              Secondary / secondary ~ Married
##  3                NA Pensioner            Secondary / secondary ~ Married
##  4                NA Pensioner            Higher education        Separated
##  5                NA Working              Secondary / secondary ~ Single / not ~
##  6                NA Commercial associate Higher education        Single / not ~
##  7                NA Pensioner            Secondary / secondary ~ Married
##  8                NA Commercial associate Higher education        Married
##  9                NA Working              Secondary / secondary ~ Married
## 10                NA Commercial associate Secondary / secondary ~ Married
## # i 13 more rows
```

3.3 Calculate the Interquartile Range for Annual_income_usd using the function IQR, remove any NA. Create an object to save the result IQR_AI, print the result.

```
IQR_AI <- IQR(CCC$Annual_income_usd, na.rm = TRUE)
IQR_AI
```

```
## [1] 103500
```

What is the IQR? Explain in your own words

The Interquartile Range (IQR) measures the spread of the middle 50% of data and is calculated as the difference between the third and first quartiles, giving a good measure of variability by showing the range within which the central half of the data lies.

3.4 One popular technique to deal with outliers is to replace them with the mean or median of the variable. For the variable Annual_income_usd, replace any observation above Quartile3 + 1.5*IQR with the mean, create a new variable to do that: Annual_income_usd_mean. You can use the function quantile to get the quartiles, save Quartile 3 into an object Q3_AI.

```
quantile(CCC$Annual_income_usd, na.rm = TRUE)
```

```
##      0%     25%     50%     75%    100%
##   33750  121500  166500  225000 1575000
```

```
Q3_AI <- 225000
CCC %>%
  mutate(Annual_income_usd_mean = replace(Annual_income_usd, Annual_income_usd > Q3_AI + 1.5 * IQR_AI, 
  drop_na() %>%
  select(Annual_income_usd, Annual_income_usd_mean)
```

```
## # A tibble: 1,525 x 2
##    Annual_income_usd Annual_income_usd_mean
##                <dbl>                  <dbl>
##  1            180000                 180000
##  2            315000                 315000
##  3            315000                 315000
##  4            315000                 315000
##  5            315000                 315000
##  6            315000                 315000
##  7            180000                 180000
##  8            450000                 191399.
##  9            450000                 191399.
## 10            450000                 191399.
## # i 1,515 more rows
```

3.5 Another popular technique is to remove the outliers. It is better to create a safety copy of the dataset instead of changing the original one. Create a safety copy for CCC: CCC_deleted, for this, remove the outliers.Show the header of the new object.

```
CCC_deleted <- CCC[-c(out_index), ]
head(CCC_deleted)
```

```
## # A tibble: 6 x 4
##   Annual_income_usd Type_Income          EDUCATION        Marital_status
##               <dbl> <chr>                <chr>            <chr>
## 1            180000 Pensioner            Higher education Married
## 2            315000 Commercial associate Higher education Married
## 3            315000 Commercial associate Higher education Married
## 4                NA Commercial associate Higher education Married
## 5            315000 Commercial associate Higher education Married
## 6            315000 Pensioner            Higher education Married
```

3.6 Now, determine how many NA's you have for Annual_income_usd, then, drop them and create a new dataset: CCC_deleted_complete.

```
sum(is.na(CCC_deleted$Annual_income_usd))
```

```
## [1] 23
```

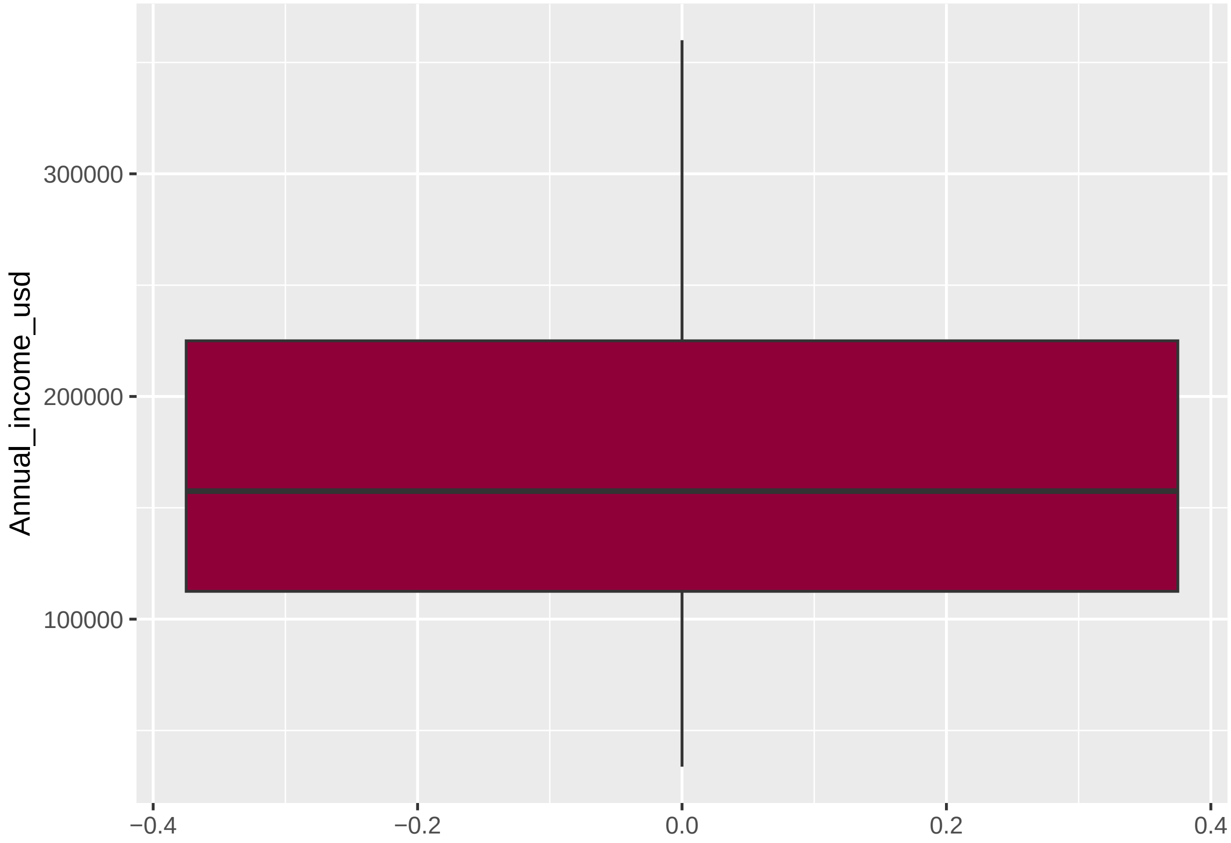
```
CCC_deleted_complete <- CCC_deleted %>%
  drop_na(Annual_income_usd)

head(CCC_deleted_complete)
```

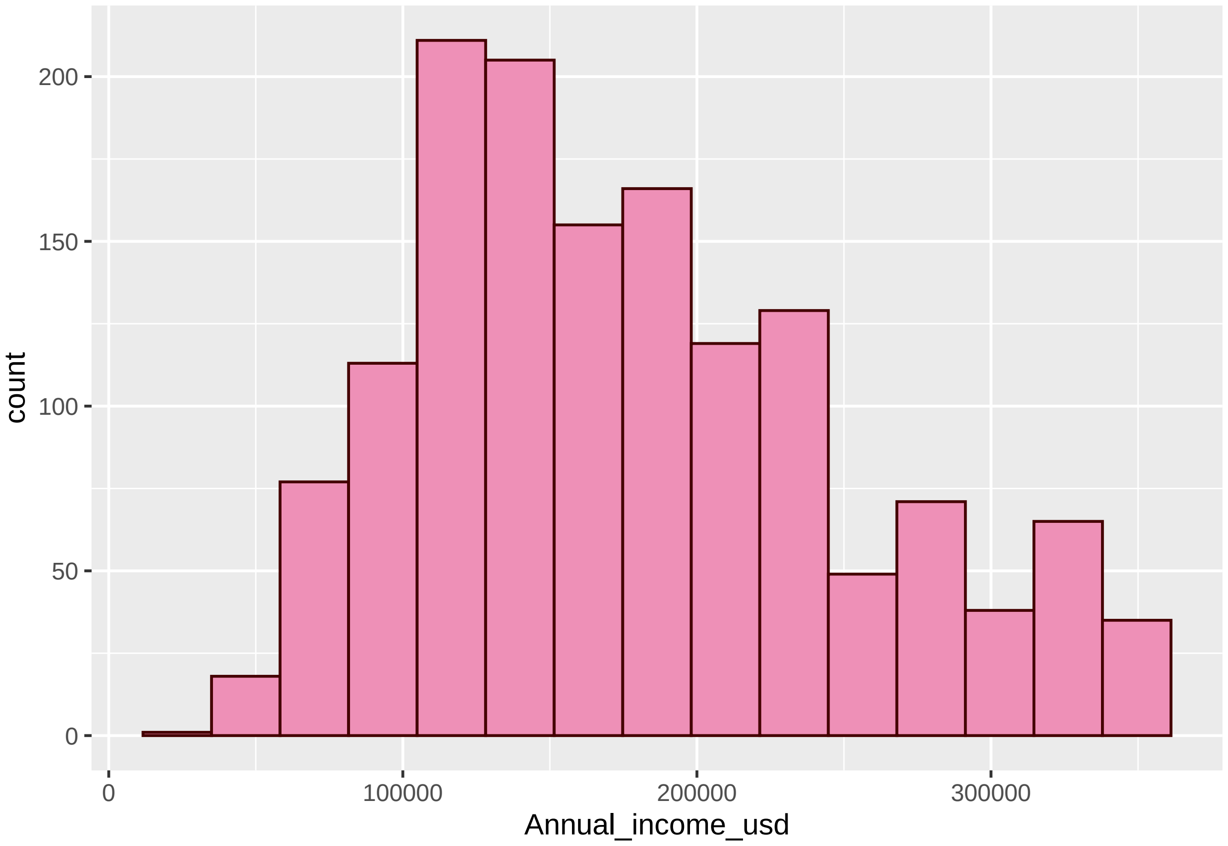
```
## # A tibble: 6 x 4
##   Annual_income_usd Type_Income          EDUCATION        Marital_status
##               <dbl> <chr>                <chr>            <chr>
## 1            180000 Pensioner            Higher education Married
## 2            315000 Commercial associate Higher education Married
## 3            315000 Commercial associate Higher education Married
## 4            315000 Commercial associate Higher education Married
## 5            315000 Pensioner            Higher education Married
## 6            315000 Commercial associate Higher education Married
```

3.7 Now, create a new boxplot, histogram (with 15 bins) and density plot that shows a line for the mean value, all for annual_income_usd variable.

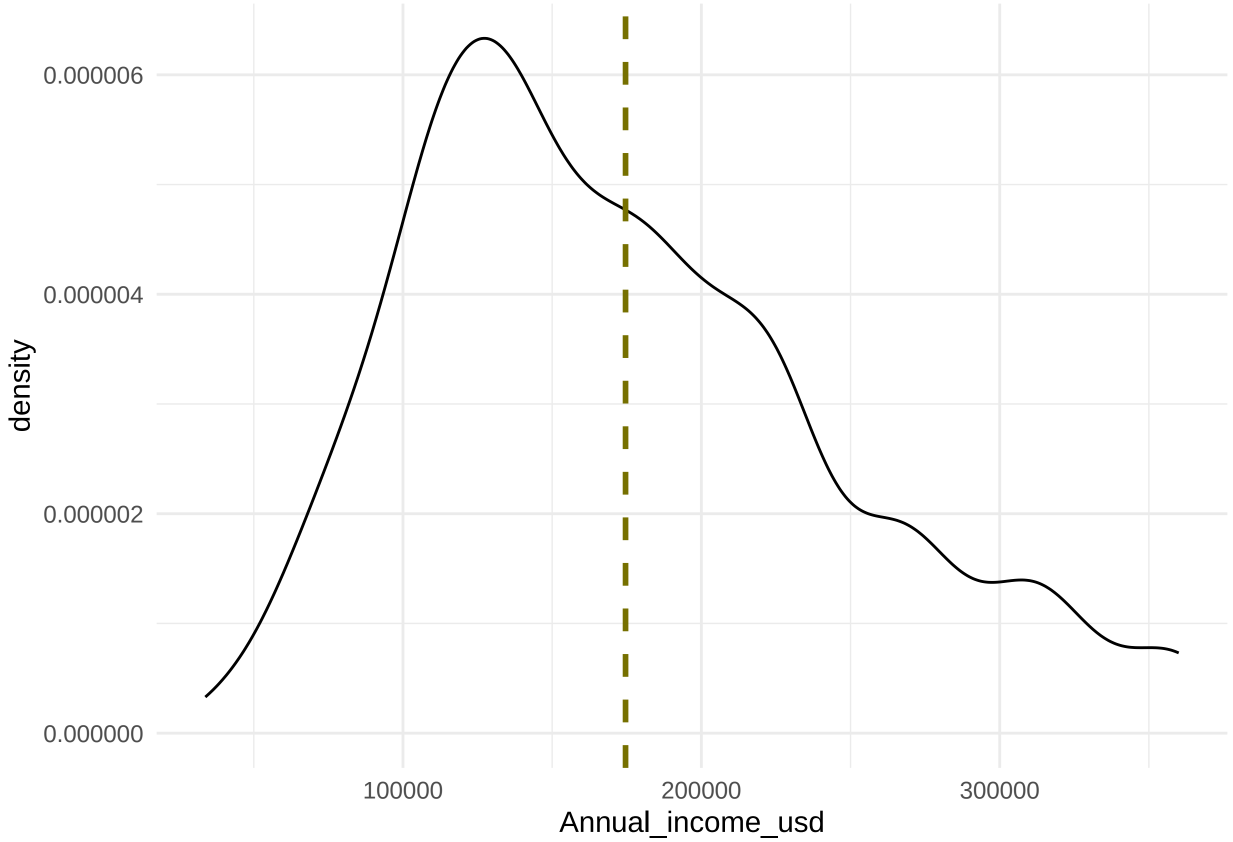
```
options(scipen = 999) # This will avoid scientific notation in the numbers of the graph
CCC_deleted_complete %>%
  ggplot(aes(y = Annual_income_usd)) +
  geom_boxplot(fill = "#8f0039")
```



```
CCC_deleted_complete %>%
  ggplot(aes(x = Annual_income_usd)) +
  geom_histogram(color = "#440000", fill = "#ee90b7", bins = 15)
```



```
CCC_deleted_complete %>%
  ggplot(aes(x = Annual_income_usd)) +
  geom_density() +
  geom_vline(aes(xintercept = mean(Annual_income_usd)), color = "#767000", linetype = "dashed", size =
  theme_minimal()
```



Part IV. Modeling the relationship between three variables

4.1 Prepare a table (with margins) that shows the relationship between the variable Marital Status Vs Education Level, that is, this table should describe the amount of people inside each category, for instance, how many people with higher education are married, single, etc. Add a mosaic plot to present the information.

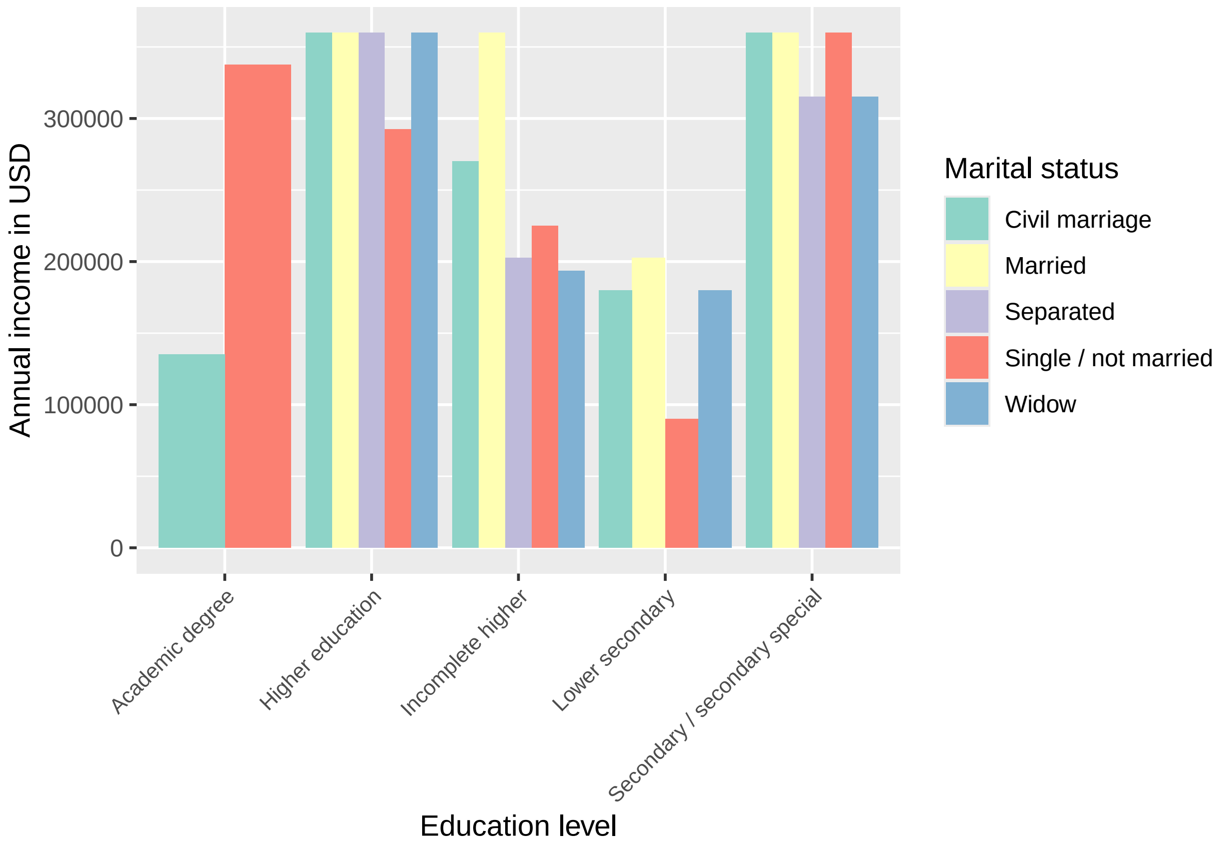
```
Ed_Vs_Marital <- table(CCC$EDUCATION, CCC$Marital_status) %>%
  addmargins()
Ed_Vs_Marital
```

```
##
##                                   Civil marriage Married Separated
##   Academic degree                              1       0         0
##   Higher education                            18     309        34
##   Incomplete higher                           10      39         5
##   Lower secondary                              2      14         0
##   Secondary / secondary special               70     687        57
##   Sum                                        101    1049        96
##
##                                   Single / not married Widow  Sum
##   Academic degree                                    1     0    2
##   Higher education                                  59     6  426
##   Incomplete higher                                 11     3   68
##   Lower secondary                                    3     2   21
##    Secondary / secondary special                      153     64 1031
##    Sum                                                227     75 1548
```

4.2 Finally, create a graph that shows in the x axis, the education level, in the y axis, the annual income in usd, and it is filled with the marital status variable.

```r
Ed_Vs_Marital_df <- data.frame(Ed_Vs_Marital)

ggplot(CCC_deleted_complete, aes(
  fill = Marital_status, x = EDUCATION, y = Annual_income_usd
)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme(axis.text.x = element_text(size = rel(0.9), angle = 45, hjust = 1)) +
  labs(
    x = "Education level", y = "Annual income in USD", fill = "Marital status"
  ) +
  scale_fill_brewer(palette = "Set3")
```



Describe what you see in this graph

- Academic degree: The lowest income for civil marriages and the highest for those who are single or not married. The plot doesn't show anything about the others.
- Higher education: Generally high incomes across all marital statuses, with single/not married having the lowest incomes.
- Incomplete higher: Moderate to high incomes, with civil marriage and married individuals showing the highest ones.

- Lower secondary: Lower incomes compared to the others levels, with single/not married showing the lowest.
- Secondary: The highest incomes in the plot.