Project 1 - Travel Insurance Data

Kevin Lopez, Clarisse Sarah, and Mayank Thacker

2024-02-18

# Work break down:

### Kevin

* Create shared Rmd file and combined code from team members. [GitHub Repository](https://github.com/KevinDLopez/CECS450-Project1)
* Created Report using Rmd ( Not much work, Started Rmd to be well formated for word).
* Created Graph 4 and Graph 5 6.

### Clarisse

* Created Graph 1, 2.
* Worked on PPT.

### Mayank

* Created Graph 3.
* Created Worked on PPT

# Explain the dataset

The dataset is a travel insurance dataset. It contains information travelers and whether they purchased travel insurance or not and other information (age, income, Employment, etc). The dataset contains 10 columns and 1986 rows of information. - <https://www.kaggle.com/datasets/tejashvi14/travel-insurance-prediction-data>

### The columns are:

**Age**- Age Of The Customer

**Employment Type**- The Sector In Which Customer Is Employed

**GraduateOrNot**- Whether The Customer Is College Graduate Or Not

**AnnualIncome**- The Yearly Income Of The Customer In Indian Rupees[Rounded To Nearest 50 Thousand Rupees]

**FamilyMembers**- Number Of Members In Customer’s Family

**ChronicDisease**- Whether The Customer Suffers From Any Major Disease Or Conditions Like Diabetes/High BP or Asthama,etc.

**FrequentFlyer**- Derived Data Based On Customer’s History Of Booking Air Tickets On Atleast 4 Different Instances In The Last 2 Years[2017-2019].

**EverTravelledAbroad**- Has The Customer Ever Travelled To A Foreign Country[Not Necessarily Using The Company’s Services]

**TravelInsurance**- Did The Customer Buy Travel Insurance Package During Introductory Offering Held In The Year 2019.

# Load the Dataset

data = read.csv("./TravelInsurancePrediction.csv")  
#knitr::kable(head(data), caption = "Table 1: First 6 Rows of the Dataset")  
# colnames(data)  
library(flextable)  
library(ggplot2)  
library(reshape2)  
  
display\_table = function(data){  
 # Function to display a nice looking table using flextable.   
 ## It applies a minimal theme (Shows the vertical and horizontal lines of the table).  
 ## https://ardata-fr.github.io/flextable-book/  
 my\_table <- head(data)   
 table = flextable(my\_table)  
 table = fontsize(table, size = 7, part = "all")  
 table = align(table, align = "center", part = "all")  
 table = theme\_box(table) # Apply a predefined theme ( Simple table )  
 table = align(table, align = "center", part = "all")  
 table = bold(table, part = "header") # Bold text in header  
 table = width(table, width = .7)  
 table = width(table, j = 1, width = .2) # Adjust width of the first column  
 return(table) # Display the flextable  
}  
  
display\_table(data)

| **X** | **Age** | **Employment.Type** | **GraduateOrNot** | **AnnualIncome** | **FamilyMembers** | **ChronicDiseases** | **FrequentFlyer** | **EverTravelledAbroad** | **TravelInsurance** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 31 | Government Sector | Yes | 400,000 | 6 | 1 | No | No | 0 |
| 1 | 31 | Private Sector/Self Employed | Yes | 1,250,000 | 7 | 0 | No | No | 0 |
| 2 | 34 | Private Sector/Self Employed | Yes | 500,000 | 4 | 1 | No | No | 1 |
| 3 | 28 | Private Sector/Self Employed | Yes | 700,000 | 3 | 1 | No | No | 0 |
| 4 | 28 | Private Sector/Self Employed | Yes | 700,000 | 8 | 1 | Yes | No | 0 |
| 5 | 25 | Private Sector/Self Employed | No | 1,150,000 | 4 | 0 | No | No | 0 |

## Reshaping data

Data is already in long format. Data for income is given in Indian Rupees. We will create a new column to convert it to USD. Column name will be AnnualIncomeUSD

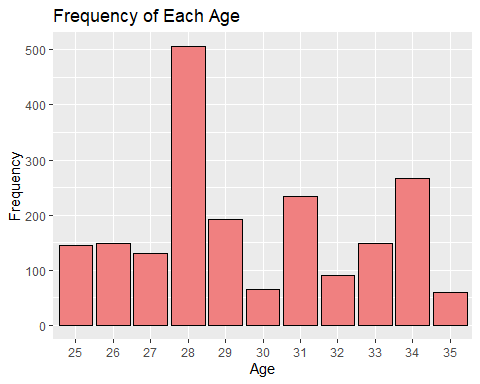
RUPEE\_TO\_USD = 0.012  
# add a new column for income in USD  
data$AnnualIncomeUSD <- data$AnnualIncome \* RUPEE\_TO\_USD  
table = display\_table(data)  
  
# hihglight the last column foreground  
table = color(table, color = "orange", part = "all", j = 11)  
table

| **X** | **Age** | **Employment.Type** | **GraduateOrNot** | **AnnualIncome** | **FamilyMembers** | **ChronicDiseases** | **FrequentFlyer** | **EverTravelledAbroad** | **TravelInsurance** | **AnnualIncomeUSD** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 31 | Government Sector | Yes | 400,000 | 6 | 1 | No | No | 0 | 4,800 |
| 1 | 31 | Private Sector/Self Employed | Yes | 1,250,000 | 7 | 0 | No | No | 0 | 15,000 |
| 2 | 34 | Private Sector/Self Employed | Yes | 500,000 | 4 | 1 | No | No | 1 | 6,000 |
| 3 | 28 | Private Sector/Self Employed | Yes | 700,000 | 3 | 1 | No | No | 0 | 8,400 |
| 4 | 28 | Private Sector/Self Employed | Yes | 700,000 | 8 | 1 | Yes | No | 0 | 8,400 |
| 5 | 25 | Private Sector/Self Employed | No | 1,150,000 | 4 | 0 | No | No | 0 | 13,800 |

# Graph 1 - Frequency of Travel Insurance

* we still need to add info here!

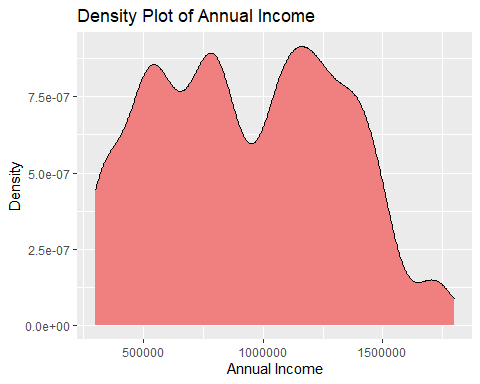
# Graph 4 - Histogram of Age  
ggplot(data, aes(x = factor(Age))) +  
 geom\_bar(fill = "lightcoral", color = "black") + # create histogram based on Age.   
 labs(title = "Frequency of Each Age", x = "Age", y = "Frequency") # add title and labels for x,y axis



# Graph 2 - Density Plot of Income

* we still need to add info here!

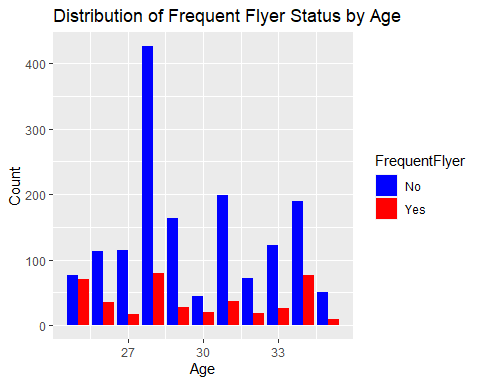
ggplot(data, aes(x = AnnualIncome)) +  
 geom\_density(fill = "lightcoral", color = "black") +  
 labs(title = "Density Plot of Annual Income", x = "Annual Income", y = "Density")



# Graph 3 - Bar Plot of Frequent Flyer by age

* we still need to add info here!

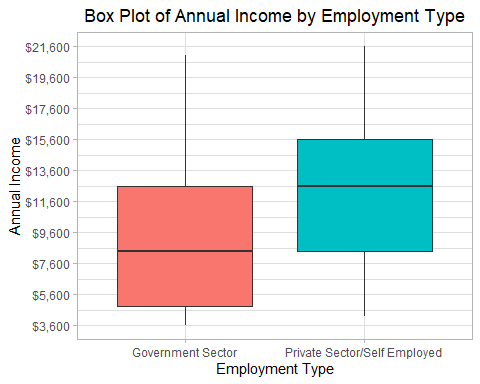
# grouped bar chart  
frequentFlyerStatus <- c("Yes" = "red", "No" = "blue")  
ggplot(data, aes(x = Age, fill = FrequentFlyer)) +  
 geom\_bar(position = "dodge") +  
 scale\_fill\_manual(values = frequentFlyerStatus) +  
 labs(title = "Distribution of Frequent Flyer Status by Age", x = "Age", y = "Count")



# Graph 4 - Sector/Income

* We can see how the income is distributed by Employment Type (Goverment vs Private).
* We can see that the income is slightly higher in the private sector.

ggplot(data, aes(x = Employment.Type, y = AnnualIncomeUSD, fill = Employment.Type)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of Annual Income by Employment Type", x = "Employment Type", y = "Annual Income") + # Sets the title and label for x,y axis  
 scale\_y\_continuous(labels = scales::dollar\_format(), breaks = seq(from = min(data$AnnualIncomeUSD), to = max(data$AnnualIncomeUSD), by = 2000)) + # Formats the y-axis labels as dollars and sets breaks  
 theme\_light() +   
 theme(   
 legend.position = "none", # Hides the legend   
 plot.title = element\_text(hjust = 0.5), # Center the plot title  
 )



# Graph 5 - Age/Income

Age and income is a cool graph to see. We can see how income is distributed by age. - Generally it is believed that income increases with age. - Peak around 40-50 since that is when people are at their peak in their careers.

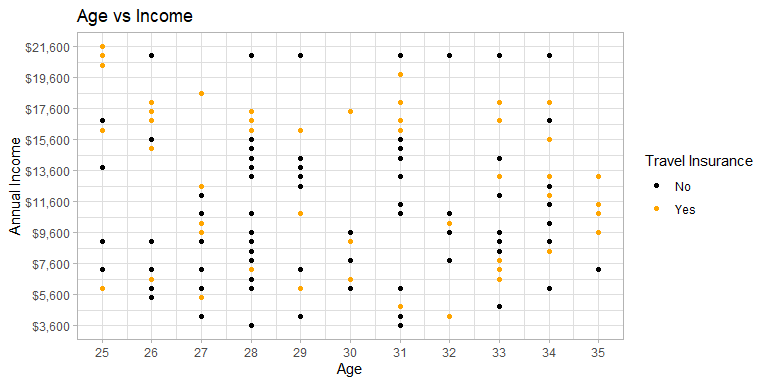
## What does the data tell us

* There is not a big difference in the age and the income that these people make.

## By looking at a barplot for each age,

* There is a slight increase of people who buy insurance as they make more money.
* It is easier to tell in box plot compared to scatter plot.

p = ggplot(data, aes(x = Age, y = AnnualIncomeUSD)) +  
 geom\_point(aes(color = as.factor(TravelInsurance))) + # Convert TravelInsurance to a factor for color mapping  
 scale\_x\_continuous(breaks = seq(min(data$Age), max(data$Age), by = 1)) + # Sets the x-axis breaks to increment by 1 (Age)   
 scale\_y\_continuous(labels = scales::dollar\_format(), breaks = seq(from = min(data$AnnualIncomeUSD), to = max(data$AnnualIncomeUSD), by = 2000)) + # Formats the y-axis labels as dollars and increment by 2k   
 scale\_color\_manual(values = c("0" = "black", "1" = "orange"), name = "Travel Insurance", labels=c("No ", "Yes")) + # Sets custom colors for TravelInsurance and customizes legend title  
 labs(title = "Age vs Income", x = "Age", y = "Annual Income") + # Sets the title and label for x,y axis   
 theme\_light()  
p



## Age/Income as box-plot

ggplot(data, aes(x = as.factor(Age), y = AnnualIncomeUSD, fill=as.factor(TravelInsurance) )) + # Plot Age ( min-max) vs AnnualIncomeUSD with fill color as TravelInsurance  
 geom\_boxplot() +  
 labs(title = "Box Plot of Annual Income by Age", x = "Age", y = "Annual Income", fill = "Travel\nInsurance" ) + # Sets the title and label for x,y axis and legend title  
 scale\_y\_continuous(labels = scales::dollar\_format(), breaks = seq(from = min(data$AnnualIncomeUSD), to = max(data$AnnualIncomeUSD), by = 2000)) + # Formats the y-axis labels as dollars and increase by 2k   
 # change the text of legend title   
 scale\_fill\_manual(values = c("0" = "gray", "1" = "green"), labels = c("No", "Yes")) + # Sets custom colors and labels for the legend  
 theme\_light() +   
 theme( plot.title = element\_text(hjust = 0.5) ) # Centers the plot title

