# Project 1 - Travel Insurance Data

Kevin Lopez, Clarisse Bonang, and Mayank Thacker

2024-02-18

## Work breakdown:

### Kevin Lopez

- Create shared Rmd file and combined code from team members. GitHub Repository
- Created Report using Rmd ( Not much work, Started Rmd to be well formated for word).
- Created Graph 4 and Graph 5,6.

### Clarisse Bonang

- Created Graph 3. And description/analysis of other graphs.
- Worked on PPT, and report.

### Mayank Thacker

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

## Explain the dataset

The travel insurance dataset provides valuable information about travelers, offering insights into their travel insurance purchasing decisions. In addition, the dataset includes details such as age, income, employment status, and more to highlight factors influencing insurance choices. With 10 columns and 1986 rows, the data offers robust information for conducting insightful analyses and patterns within the dataset.

- The data was collected at Guru Gobind Singh Indraprastha University (GGSIPU).
- To access the dataset please visit kaggle.
- The dataset was collected from India.

### 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 |

| X | Age | Employment.Type | GraduateOrNot | AnnualIncome | FamilyMembers | ChronicDiseases | FrequentFlyer | EverTravelledAbroad | TravelInsurance |
|---|---|---|---|---|---|---|---|---|---|
| 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
```

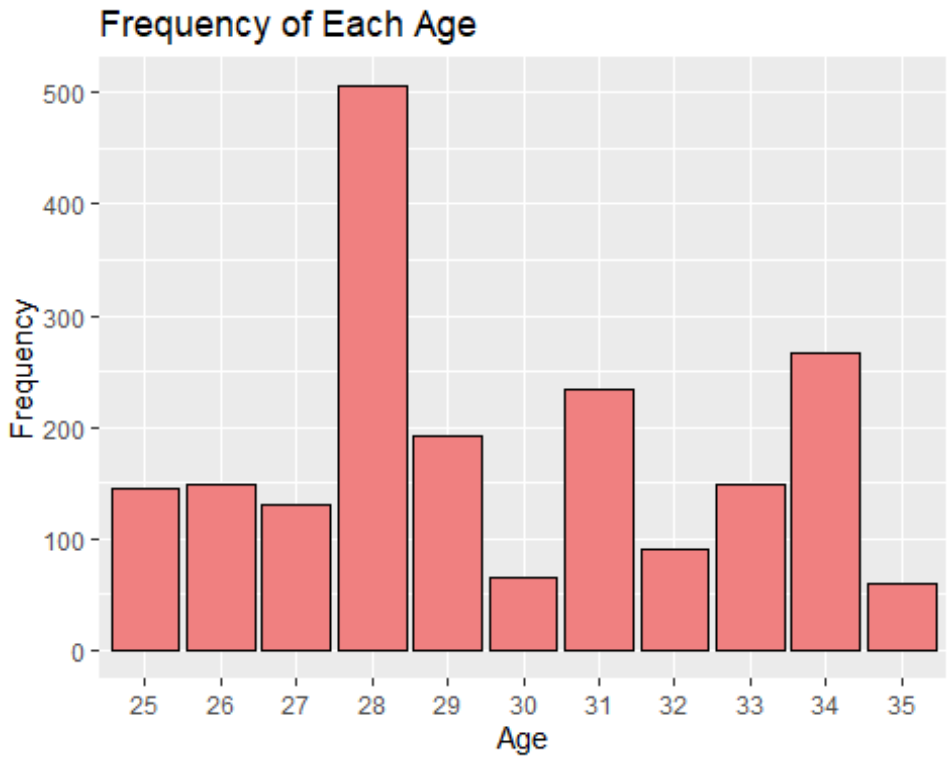
| X | Age | Employment.Type | GraduateOrNot | AnnualIncome | FamilyMembers | ChronicDiseases | FrequentFlyer | EverTravelledAbroad | TravelInsurance | AnnualIncomeUSD |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 31 | Government Sector | Yes | 400,000 | 6 | 1 | No | No | 0 | 4,800 |
| 1 | 31 | 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 Age Distribution

The histogram represents the frequency distribution of individual ages in the dataset. It provides insights into the prevalence of various age groups within the dataset. For instance, 28 years old has the highest frequency, followed by 34 years old. Beyond these notable frequencies, there is an even distribution across the remaining age categories, highlighting a balanced representation amongst the age values.

- The histogram shows the frequency of each individual age in the dataset. It provides insights into the prevalence of different age groups within the dataset.
- 28 years old has the highest frequency

```
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
```



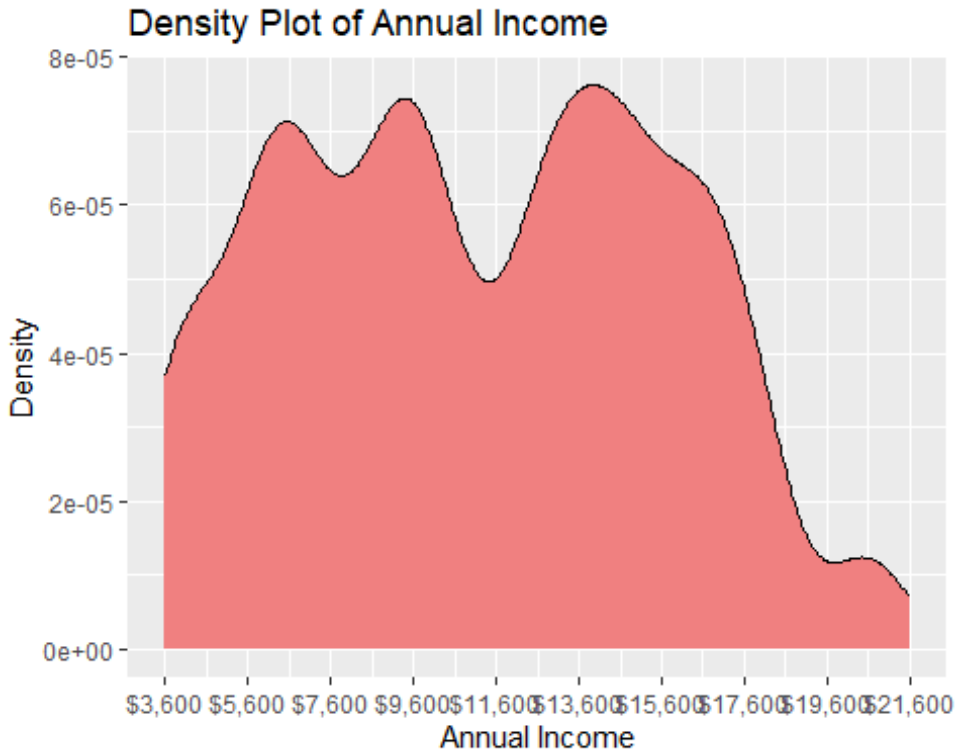

## Graph 2 - Annual Income Density

The graph represents the probability density distribution across a range of annual income values, constrained between a minimum of $3600 and a maximum of $21600. Peaks in probability density occur at $5600, $9600, and $13600. However, the graph highlights a dip in probability density between $9600 to $11600 and $17600 to $19600. Based on these probabilities, we anticipate a higher frequency of annual incomes at $5600, $9600, and

$13600, indicating a slight disparity within our population pool concerning annual income distributions.

- The graph shows the probability density of Annual Income.
- X-axis: “Annual Income”, Y-axis: “Density”.

```
ggplot(data, aes(x = AnnualIncomeUSD)) + # Set the x-axis to AnnualIncome
  geom_density(fill = "lightcoral", color = "black") + # Add a density layer
with lightcoral fill and black border
  labs(title = "Density Plot of Annual Income", x = "Annual Income", y =
"Density") + # Add title and axis labels
  scale_x_continuous(labels = scales::dollar_format(), breaks = seq(from =
min(data$AnnualIncomeUSD), to = max(data$AnnualIncomeUSD), by = 2000)) # add
$ and increments by 2000
```

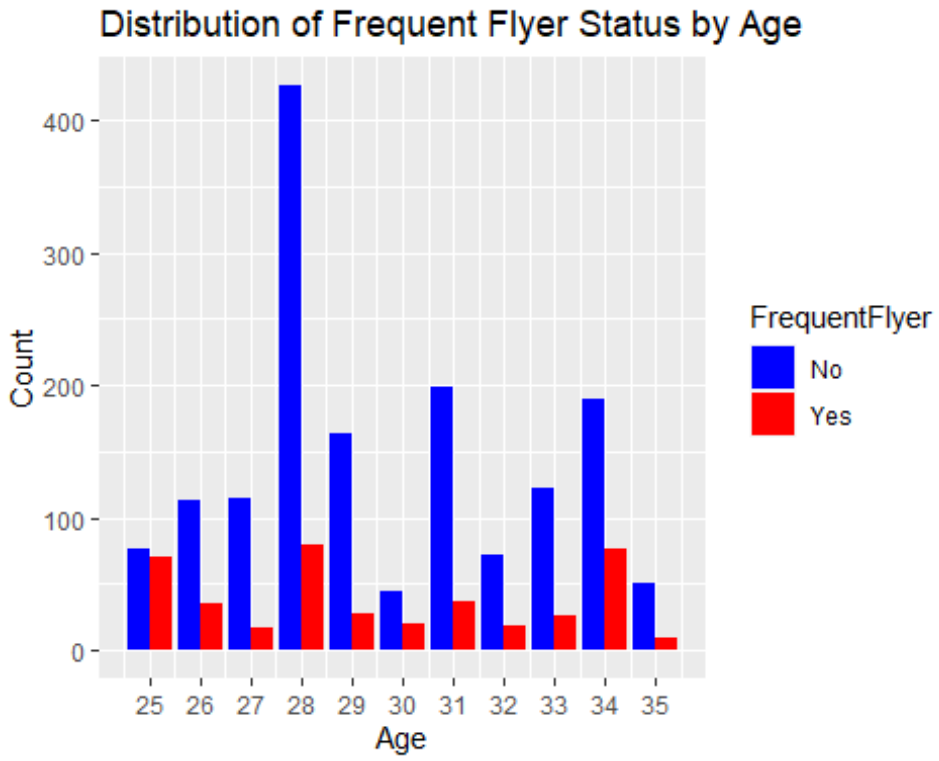



## Graph 3 - Bar Plot of Frequent Flyer by age

- The graph highlights the distribution of flyer status across different age groups.
- It’s evident that there is a higher number of non-frequent flyers than that of frequent flyers between the ages of 25 to 35.
- In particular, there is a decline in the number of frequent flyers between ages 25 to 27 and 28 to 30 and a steady increase in non-frequent flyers. In comparison, between ages 32 to 24, there is an increase in the number of frequent flyers and non-frequent flyers.

- Notably, 28 year-olds appear to be less likely to be frequent flyers based on the lower count from the graph.
- In conclusion, the data suggests that the proportion of frequent and non-frequent flyers varies across different age groups.

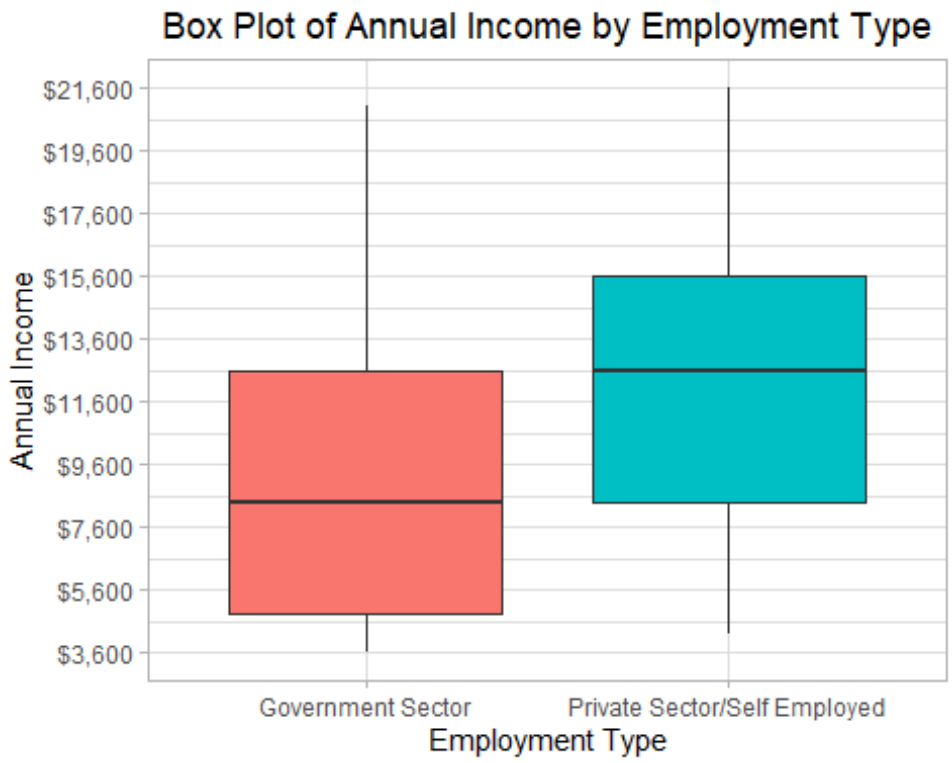
```
# grouped bar chart
frequentFlyerStatus <- c("Yes" = "red", "No" = "blue")
ggplot(data, aes(x = Age, fill = FrequentFlyer)) +
  geom_bar(position = "dodge") + #Ensures bars for frequent and non-frequent flyers are side by side
  scale_fill_manual(values = frequentFlyerStatus) + #Fills colors for both flyer status values
  labs(title = "Distribution of Frequent Flyer Status by Age", x = "Age", y = "Count") +
  scale_x_continuous(breaks = seq(25, 35, 1))  # Adding ticks for ages 25 to 35
```




## Graph 4 - Sector/Income

- The following graph shows how the income is distributed by Employment type ( Government vs Private).
- The Graph shows that interquartille range income is slightly higher in the private sector.
- Income data does not contain any outliers.

```r
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

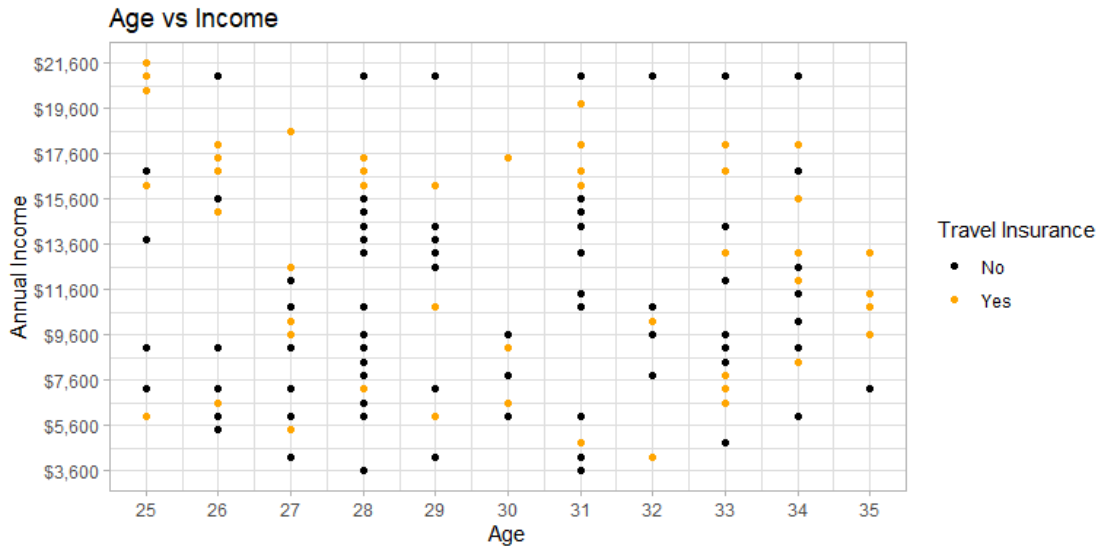
The graph shows income related to age. And see the income distribution by age.

- Generally, it is believed that income increases with age.
  - Income peak around 40-50 years old, since that is when an individual will be at the peak of their careers.

## What does the data tell us?

- This data does not fit the idea that income increases as age increases.
- There is a slight increase in people who buy insurance as they make more money.
  - It is easier to tell in boxplot 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()
# print(min(data$AnnualIncomeUSD) )
# print(max(data$AnnualIncomeUSD))
p
```




## Age/Income as box-plot

- This graph contains the same data plotted as the graph before ( Annual Income v Age).
- This plot shows that the median income(and Interquartile) is higher for those that purchase travel insurance.

```
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" = "orange"), 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
```