Travel Assured analysis

Michal Lauer

27.10.2022

Preface

This analysis is done on generated data and does not represent a real company. The main purpose of this work was to get my certification, which I've accomplished. Now I use this work to showcase my work and how I *do* data analysis. The project is located on my Github page

Introduction

Travel Assured is a travel insurance company. Due to the COVID pandemic, they have had to cut their marketing budget by over 50%. It is more important than ever that they advertise in the right places and to the right people.

Travel Assured wants answers to two key questions:

1) Are there differences in the travel habits between customers and non-customers?
2) What is the typical profile of customers and non-customers?

This analysis notebook's purpose is to introduce the analysis to a *data* person. It shows how the data was manipulated, what was changed, and how things were computed. For a more business-detail approach, see the attached presentation.

Libraries

This one block loads all libraries.

```
# Data wrangling
library(dplyr)
library(tidyr)
library(glue)

# Graphs
library(ggplot2)
library(patchwork)

# Tables
library(gtsummary)
```

Data load

The raw data set is loaded and transformed into a tibble.

```
data_raw <-
  read.csv(file = "input/travel_insurance.csv") |>
  as_tibble()

head(data_raw)
```

```
## # A tibble: 6 x 9
##     Age Employment.Type  Gradu~1 Annua~2 Famil~3 Chron~4 Frequ~5 EverT~6 Trave~7
##   <int> <chr>            <chr>     <int>   <int>   <int> <chr>   <chr>     <int>
## 1    31 Government Sect~ Yes      400000       6       1 No      No            0
## 2    31 Private Sector/~ Yes     1250000       7       0 No      No            0
## 3    34 Private Sector/~ Yes      500000       4       1 No      No            1
## 4    28 Private Sector/~ Yes      700000       3       1 No      No            0
## 5    28 Private Sector/~ Yes      700000       8       1 Yes     No            0
## 6    25 Private Sector/~ No      1150000       4       0 No      No            0
## # ... with abbreviated variable names 1: GraduateOrNot, 2: AnnualIncome,
## #   3: FamilyMembers, 4: ChronicDiseases, 5: FrequentFlyer,
## #   6: EverTravelledAbroad, 7: TravelInsurance
```

The column names are transformed to *snake_case* stanard using the snakecase package. Variables are then transformed, such that:

- *employment type* is now a factor
- *graduate or not* is a logical variable (true - graduated, false - did not graduate)
- *chronic disease* is now a logical variable (true - has disease, false - does not have a disease)
- *frequent flyer* is now a logical variable (true - is a frq. flr., false - is not a frq. flr.)
- *ever travelled abroad* is now a logical variable (true - has travelled, false
- has not travelled)
- *travel insurance* is now a logical variable (true - is insured, false - is not insured)

```
data <-
  data_raw |>
  rename_with(snakecase::to_snake_case) |>
  mutate(employment_type       = factor(employment_type),
         family_members        = family_members,
         graduate_or_not       = graduate_or_not == "Yes",
         chronic_diseases      = chronic_diseases == 1,
         frequent_flyer        = frequent_flyer == "Yes",
         ever_travelled_abroad = ever_travelled_abroad == "Yes",
         travel_insurance      = travel_insurance == 1)
```

The data is now translated and the overall look is shown below.

```
skimr::skim_without_charts(data)
```

Table 1: Data summary

Name	data
Number of rows	1987
Number of columns	9
Column type frequency:	
factor	1
logical	5
numeric	3
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
employment_type	0	1	FALSE	2	Pri: 1417, Gov: 570

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
graduate_or_not	0	1	0.85	TRU: 1692, FAL: 295
chronic_diseases	0	1	0.28	FAL: 1435, TRU: 552
frequent_flyer	0	1	0.21	FAL: 1570, TRU: 417
ever_travelled_abroad	0	1	0.19	FAL: 1607, TRU: 380
travel_insurance	0	1	0.36	FAL: 1277, TRU: 710

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1	29.65	2.91	25	28	29	32	35
annual_income	0	1	932762.96	376855.68	300000	600000	900000	1250000	1800000
family_members	0	1	4.75	1.61	2	4	5	6	9

Form the overview, it can be seen that there are no missing values - which is really nice! Employment has only two levels - *Government Sector*, and *Private Sector/Self Employed*. Each customer can have up to 9 other family members (excluding 1 as the factor starts at two family members). Except for customers who have bought an insurance, all logical values are heavily imbalanced. This could be useful in hypothesis testing and data exploration. Finally,

Support variables

For further data analysis, some help variables are created to reduce bugs.

```r
# Columns assigned as Travel habits
travel_habits    <- c("frequent_flyer", "ever_travelled_abroad")
# Columns assigned as Customer profiles
```

```
customer_profile <- c("age", "employment_type", "graduate_or_not",
                      "annual_income", "family_members", "chronic_diseases",
                      "travel_insurance")
# Labels transformation
labels           <- c("TRUE" = "Yes", "FALSE" = "No")
```

Difference in travel habits

The first business question that is asked is:

> Are there differences in the travel habits between customers and non-customers?

Two columns which identify travel habits are *frequent_flyer* and *ever_travelled_abroad*.

Frequent flyers

To identify if the difference among insured and uninsured people who are frequent flyers and who are **not** frequent flyers, the p-value is calculated. First the proportions are prepared in a special table.

```
fq_prop <-
  data |>
  select(frequent_flyer, travel_insurance) |>
  group_by(travel_insurance, frequent_flyer) |>
  summarise(n = n(), .groups = "drop_last") |>
  mutate(total = sum(n),
         mean = n/total)

fq_prop
```

```
## # A tibble: 4 x 5
## # Groups:   travel_insurance [2]
##   travel_insurance frequent_flyer     n total  mean
##   <lgl>            <lgl>          <int> <int> <dbl>
## 1 FALSE            FALSE           1099  1277 0.861
## 2 FALSE            TRUE             178  1277 0.139
## 3 TRUE             FALSE            471   710 0.663
## 4 TRUE             TRUE             239   710 0.337
```

Compute *p*-value where travel insurance is false.

```
x <- fq_prop$n[1:2]
n <- fq_prop$total[1:2]

fq_prop_uninsured <- prop.test(x = x, n = n, correct = F)
fq_prop_uninsured
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
```

```
## data:  x out of n
## X-squared = 1328.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.6943567 0.7480865
## sample estimates:
##    prop 1    prop 2
## 0.8606108 0.1393892
```

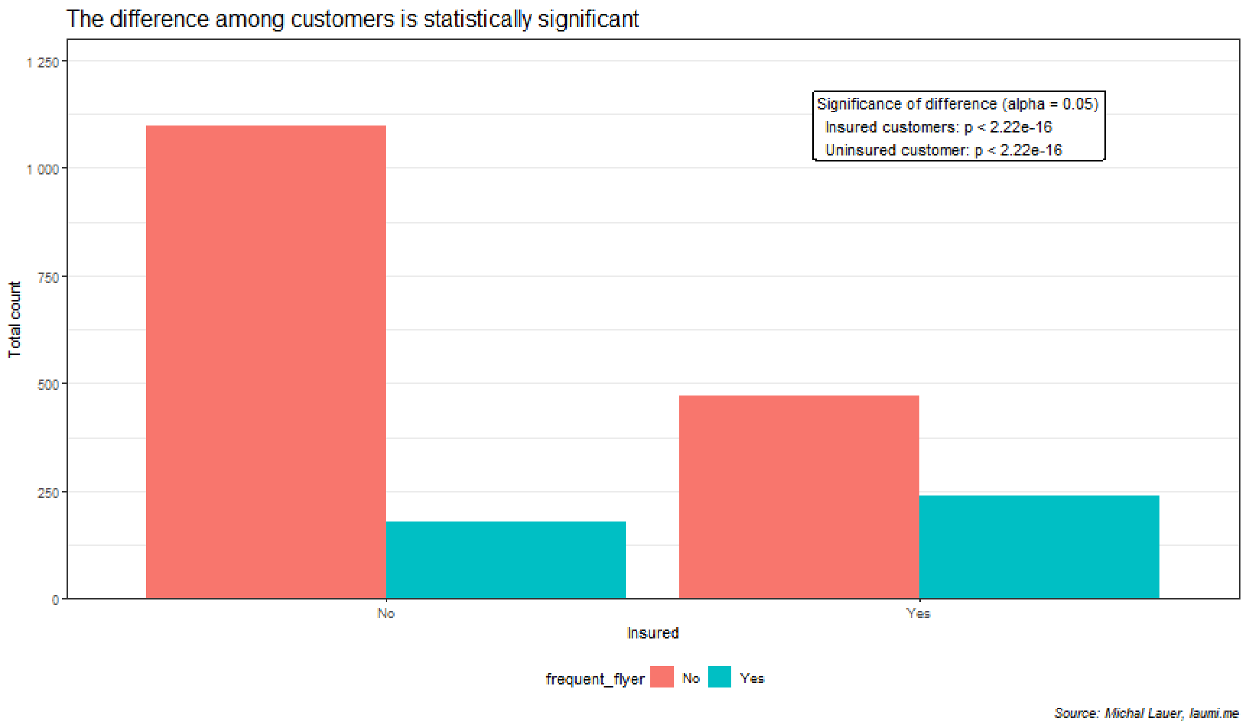
Compute p-value where travel insurance is true.

```
x <- fq_prop$n[3:4]
n <- fq_prop$total[3:4]

fq_prop_insured <- prop.test(x = x, n = n, correct = F)
fq_prop_insured
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  x out of n
## X-squared = 151.62, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2776036 0.3759175
## sample estimates:
##    prop 1    prop 2
## 0.6633803 0.3366197
```

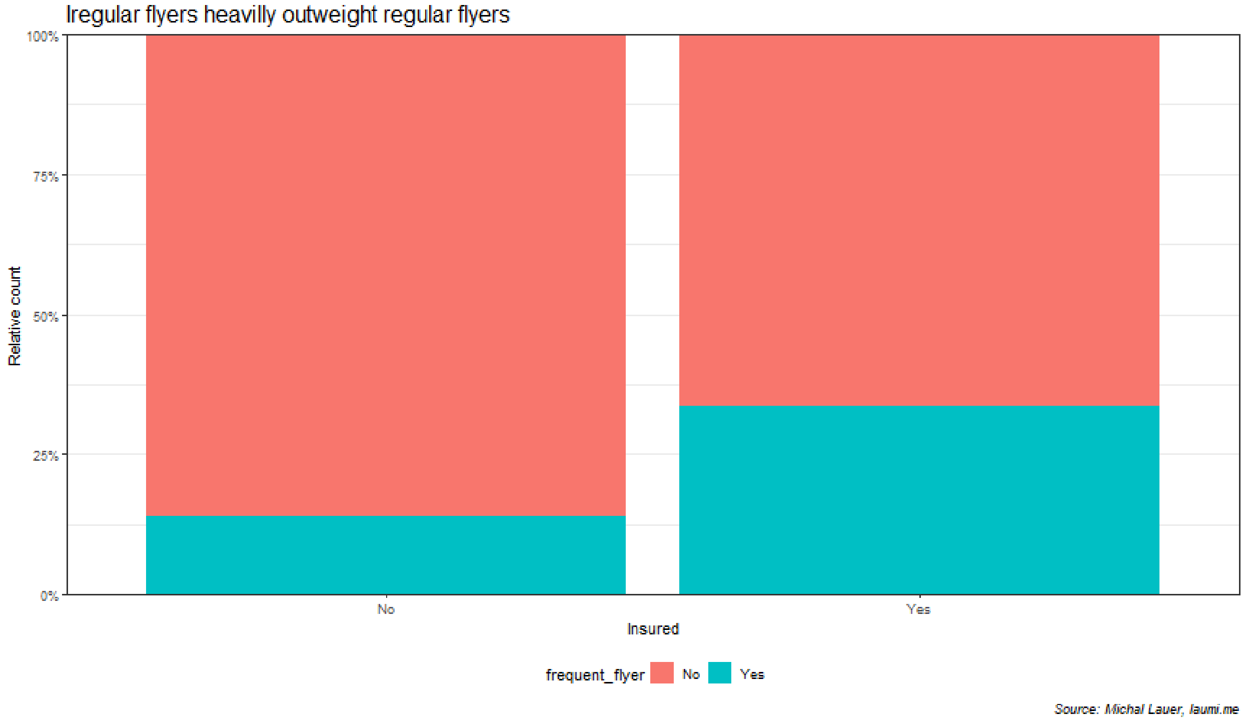
The differences among insured and uninsured people is graphically displayed.

```
# Prepare text for graph
fq_significance <- glue("
Significance of difference (alpha = 0.05)
  Insured customers: p {format.pval(fq_prop_insured$p.value)}
  Uninsured customer: p {format.pval(fq_prop_uninsured$p.value)}
")

fq_p1 <-
  data |>
  select(frequent_flyer, travel_insurance) |>
  ggplot(aes(x = travel_insurance, fill = frequent_flyer)) +
  geom_bar(position = "dodge") +
  annotate(geom = "label", x = 1.8, y = 1100,
           label = fq_significance, hjust = 0) +
  theme_bw() +
  scale_x_discrete(labels = labels) +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 1300),
                     labels = scales::label_number()) +
  scale_fill_discrete(labels = labels) +
  theme(
    panel.grid.major.x = element_blank(),
    plot.caption = element_text(face = "italic", size = 9),
    legend.position = "bottom",
    plot.title = element_text(size = 16)
  ) +
  labs(
    title = "The difference among customers is statistically significant",
    x = "Insured",
    y = "Total count",
    caption = "Source: Michal Lauer, laumi.me"
  )

fq_p1
```



The second bar plot show the proportion among insured and uninsured people.

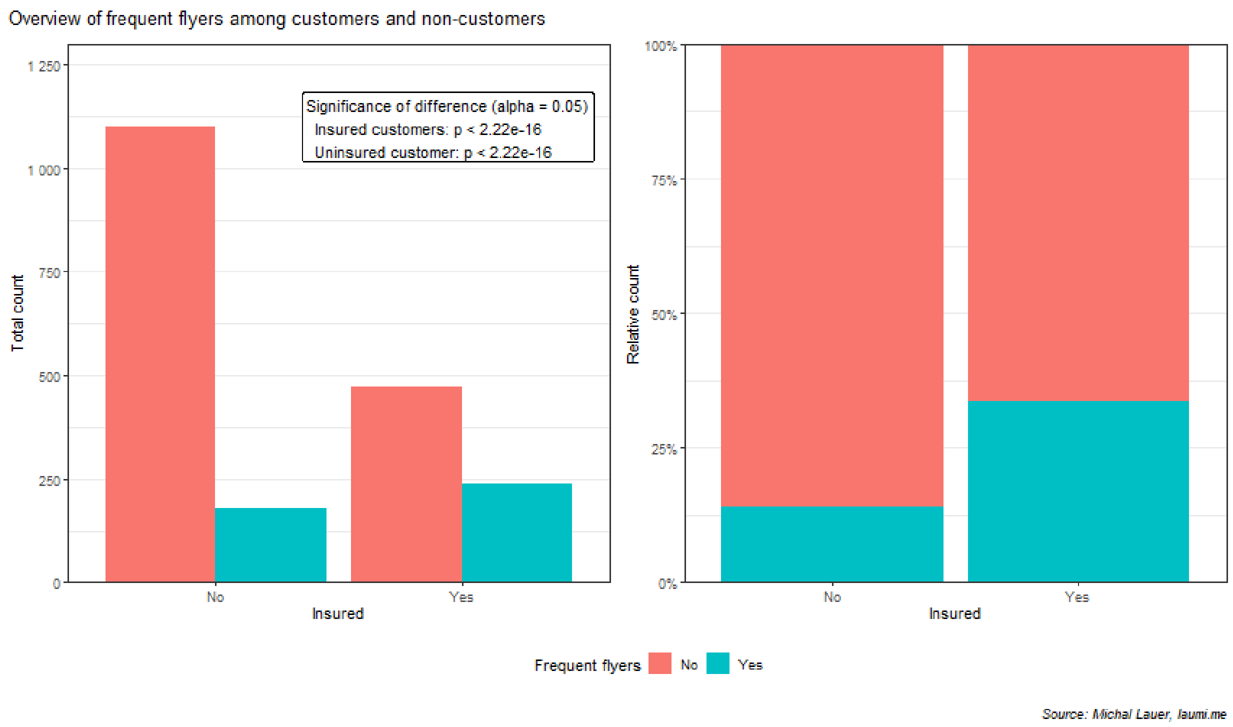
```
fq_p2 <-
  data |>
  select(frequent_flyer, travel_insurance) |>
  ggplot(aes(x = travel_insurance, fill = frequent_flyer)) +
  geom_bar(position = "fill") +
  theme_bw() +
  scale_x_discrete(labels = labels) +
  scale_y_continuous(expand = c(0, 0),
                     labels = scales::label_percent()) +
  scale_fill_discrete(labels = labels) +
  theme(
    panel.grid.major.x = element_blank(),
    plot.caption = element_text(face = "italic", size = 9),
    legend.position = "bottom",
    plot.title = element_text(size = 16)
  ) +
  labs(
    title = "Iregular flyers heavilly outweight regular flyers",
    x = "Insured",
    y = "Relative count",
    caption = "Source: Michal Lauer, laumi.me"
  )

fq_p2
```



Joined graphs using patchwork.

```
# Edit graphs for patchwork
fq_p1_pw <-
  fq_p1 +
  labs(title = NULL,
       caption = NULL)

fq_p2_pw <-
  fq_p2 +
  labs(title = NULL,
       caption = NULL)

# Update annotate position
fq_p1_pw$layers[[2]]$data$x <- 1.35

# Build graph with patchwork
```

```
fq_patchwork <-
  (fq_p1_pw + fq_p2_pw) / guide_area() +
  plot_layout(guides = "collect", heights = c(10, 1)) +
  plot_annotation(
    title = "Overview of frequent flyers among customers and non-customers",
    caption = "Source: Michal Lauer, laumi.me"
  ) & labs(
    fill = "Frequent flyers"
  ) &
  theme(plot.caption = element_text(face = "italic", size = 9))

fq_patchwork
```



Overview of frequent flyers among customers and non-customers

Source: Michal Lauer, laumi.me

Ever travelled abroad

The second identified travel habit is whether a person has ever travelled abroad. The process here is the same. First, the proportion table is created so *p*-values can be computed.

```
eta_prop <-
  data |>
  select(ever_travelled_abroad, travel_insurance) |>
  group_by(travel_insurance, ever_travelled_abroad) |>
  summarise(n = n(), .groups = "drop_last") |>
  mutate(total = sum(n),
         mean = n/total)

eta_prop
```

```
## # A tibble: 4 x 5
## # Groups:   travel_insurance [2]
##   travel_insurance ever_travelled_abroad     n total   mean
##   <lgl>            <lgl>                 <int> <int>  <dbl>
## 1 FALSE            FALSE                  1195  1277 0.936
## 2 FALSE            TRUE                     82  1277 0.0642
## 3 TRUE             FALSE                   412   710 0.580
## 4 TRUE             TRUE                    298   710 0.420
```

Compute *p*-value where travel insurance is false.

```
x <- eta_prop$n[1:2]
n <- eta_prop$total[1:2]

eta_prop_uninsured <- prop.test(x = x, n = n, correct = F)
eta_prop_uninsured
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  x out of n
## X-squared = 1940.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.8525603 0.8905877
## sample estimates:
##   prop 1   prop 2
## 0.935787 0.064213
```

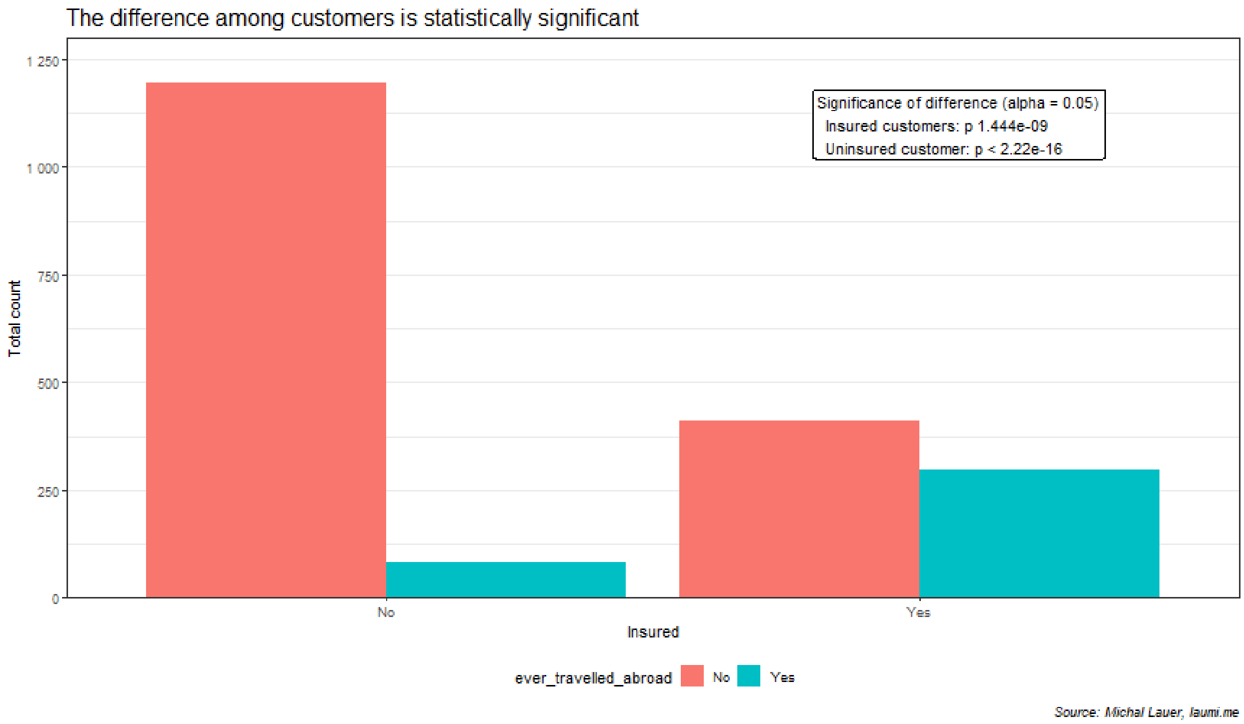
Compute *p*-value where travel insurance is true.

```
x <- eta_prop$n[3:4]
n <- eta_prop$total[3:4]

eta_prop_insured <- prop.test(x = x, n = n, correct = F)
eta_prop_insured
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  x out of n
## X-squared = 36.608, df = 1, p-value = 1.444e-09
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1092262 0.2119006
## sample estimates:
##    prop 1    prop 2
## 0.5802817 0.4197183
```

Now a bar graph is created representing the difference in each group.

```r
# Prepare text for graph
eta_significance <- glue("
Significance of difference (alpha = 0.05)
  Insured customers: p {format.pval(eta_prop_insured$p.value)}
  Uninsured customer: p {format.pval(eta_prop_uninsured$p.value)}
")

eta_p1 <-
  data |>
  select(ever_travelled_abroad, travel_insurance) |>
  ggplot(aes(x = travel_insurance, fill = ever_travelled_abroad)) +
  geom_bar(position = "dodge") +
  annotate(geom = "label", x = 1.8, y = 1100,
           label = eta_significance, hjust = 0) +
  theme_bw() +
  scale_x_discrete(labels = labels) +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 1300),
                     labels = scales::label_number()) +
  scale_fill_discrete(labels = labels) +
  theme(
    panel.grid.major.x = element_blank(),
    plot.caption = element_text(face = "italic", size = 9),
    legend.position = "bottom",
    plot.title = element_text(size = 16)
  ) +
  labs(
    title = "The difference among customers is statistically significant",
    x = "Insured",
    y = "Total count",
    caption = "Source: Michal Lauer, laumi.me"
  )

eta_p1
```

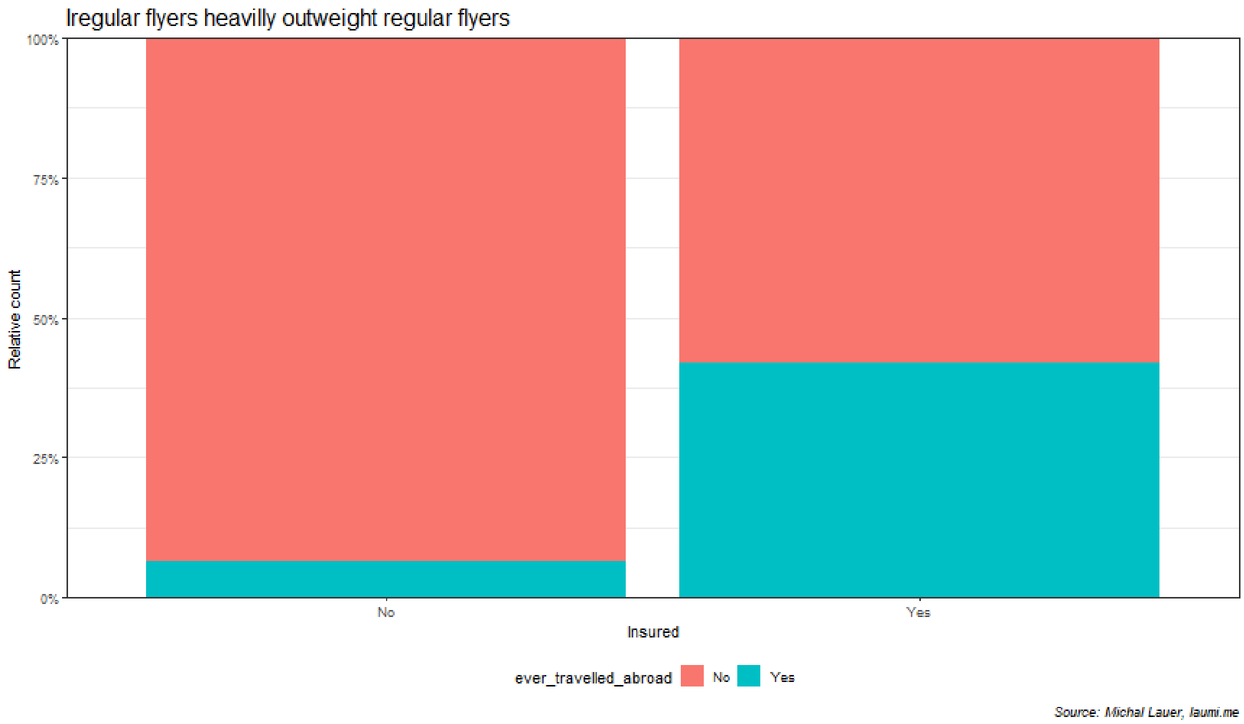


The difference among customers is statistically significant

Source: Michal Lauer, laumi.me

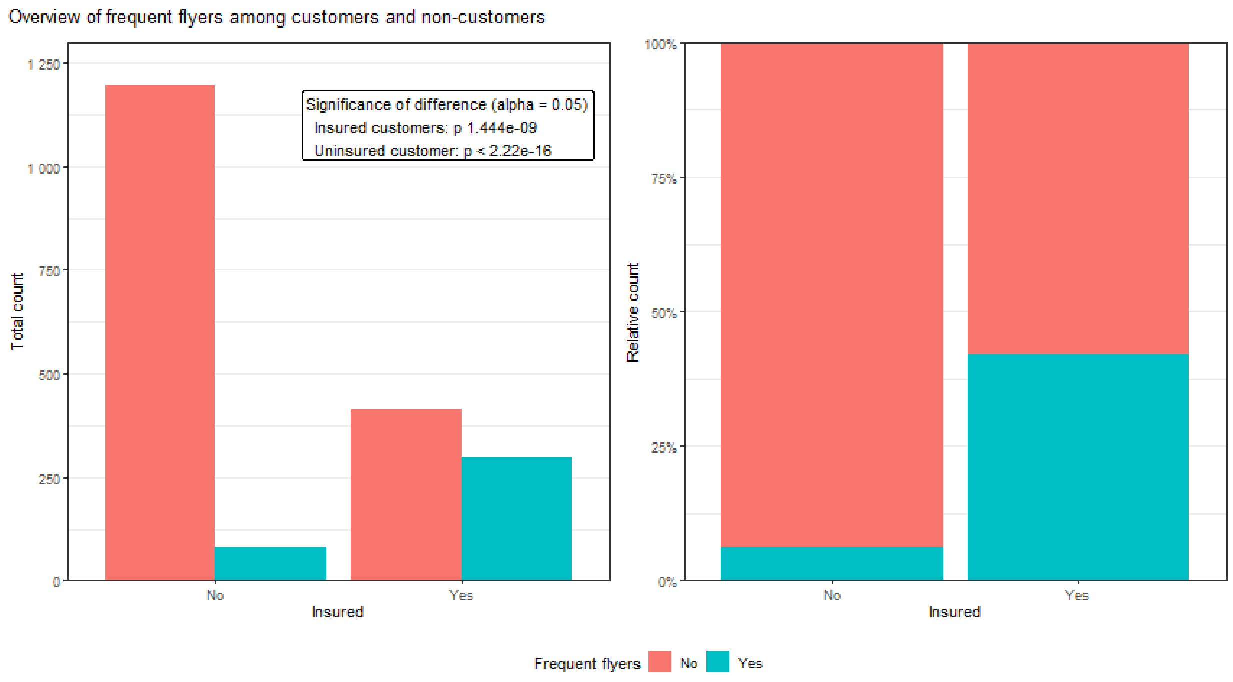
The proportions are compares using filled bar graph.

```r
eta_p2 <-
  data |>
  select(ever_travelled_abroad, travel_insurance) |>
  ggplot(aes(x = travel_insurance, fill = ever_travelled_abroad)) +
  geom_bar(position = "fill") +
  theme_bw() +
  scale_x_discrete(labels = labels) +
  scale_y_continuous(expand = c(0, 0),
                     labels = scales::label_percent()) +
  scale_fill_discrete(labels = labels) +
  theme(
    panel.grid.major.x = element_blank(),
    plot.caption = element_text(face = "italic", size = 9),
    legend.position = "bottom",
    plot.title = element_text(size = 16)
  ) +
  labs(
    title = "Iregular flyers heavilly outweight regular flyers",
    x = "Insured",
    y = "Relative count",
    caption = "Source: Michal Lauer, laumi.me"
  )

eta_p2
```



Joined graphs using patchwork.

```r
# Edit graphs for patchwork
eta_p1_pw <-
  eta_p1 +
  labs(title = NULL,
       caption = NULL)

eta_p2_pw <-
  eta_p2 +
  labs(title = NULL,
       caption = NULL)

# Update annotate position
eta_p1_pw$layers[[2]]$data$x <- 1.35

# Build graph with patchwork
eta_patchwork <-
  (eta_p1_pw + eta_p2_pw) / guide_area() +
  plot_layout(guides = "collect", heights = c(10, 1)) +
  plot_annotation(
    title = "Overview of frequent flyers among customers and non-customers",
    caption = "Source: Michal Lauer, laumi.me"
  ) & labs(
    fill = "Frequent flyers"
  ) &
  theme(plot.caption = element_text(face = "italic", size = 9))

eta_patchwork
```



Overview of frequent flyers among customers and non-customers

Source: Michal Lauer, laumi.me

Customer profile

The second questions that the business asks is:

> What is the typical profile of customers and non-customers?

This questions is answered by summarizing columns related to customer profile see here

Numerical

Overview of numerical characteristics.

```
data |>
  select(all_of(customer_profile)) |>
  select(travel_insurance, where(is.numeric)) |>
  mutate(travel_insurance = if_else(travel_insurance, "Insured", "Uninsured")) |>
  tbl_summary(by = travel_insurance,
              label = list(
                age ~ "Age",
                annual_income ~ "Annual income",
                family_members ~ "# of family members"
              ),
              statistic = everything() ~ "{mean} ± {sd} ({median})",
              digits = everything() ~ 2,
              type = family_members ~ "continuous") |>
  modify_caption("Characteristics of numerical variables") |>
  add_p()
```

```
## Table printed with ‘knitr::kable()‘, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include ‘message = FALSE‘ in code chunk header.
```

Table 5: Characteristics of numerical variables

Characteristic	**Insured**, N = 710	**Uninsured**, N = 1 277	**p-value**
Age	29.89 ± 3.33 (30.00)	29.52 ± 2.64 (29.00)	0.031
Annual income	1 133 239.44 ± 374 844.68 (1 250 000.00)	821 299.92 ± 328 898.90 (800 000.00)	<0.001
# of family members	4.93 ± 1.68 (5.00)	4.66 ± 1.56 (4.00)	<0.001

```
#TODO: Summary
```

Logical

Overview of logical characteristics.

```
data |>
  select(all_of(customer_profile)) |>
  select(travel_insurance, where(is.logical)) |>
  mutate(travel_insurance = if_else(travel_insurance, "Insured", "Uninsured")) |>
  tbl_summary(by = travel_insurance,
              label = list(
                graduate_or_not ~ "Graduated",
                chronic_diseases ~ "Has chronic disease"
              )) |>
  modify_caption("Characteristics of logical variables") |>
  add_p(pvalue_fun = function(x) style_number(x, digits = 2))
```

```
## Table printed with ‘knitr::kable()‘, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include ‘message = FALSE‘ in code chunk header.
```

Table 6: Characteristics of logical variables

Characteristic	**Insured**, N = 710	**Uninsured**, N = 1 277	**p-value**
Graduated	611 (86%)	1 081 (85%)	0.40
Has chronic disease	205 (29%)	347 (27%)	0.42

```
#TODO: Summary
```

Factor

Overview of factor characteristics.

```
data |>
  select(all_of(customer_profile)) |>
  select(travel_insurance, where(is.factor)) |>
  mutate(travel_insurance = if_else(travel_insurance, "Insured", "Uninsured")) |>
  tbl_summary(by = travel_insurance,
              label = employment_type ~ "Employment type") |>
  modify_caption("Characteristics of factor variables") |>
  add_p()
```

```
## Table printed with ‘knitr::kable()‘, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include ‘message = FALSE‘ in code chunk header.
```

Table 7: Characteristics of factor variables

Characteristic	**Insured**, N = 710	**Uninsured**, N = 1 277	**p-value**
Employment type			<0.001
Government Sector	140 (20%)	430 (34%)	
Private Sector/Self Employed	570 (80%)	847 (66%)	

```
#TODO: Summary
```