library(tidyverse)

library(patchwork)

library(readr)

## set working directory

setwd("/Users/thomas/OneDrive - INSTITUTO TECNOLOGICO AUTONOMO DE MEXICO/2016\_ITAM/2023\_MaestriaCienciasDeDatos/2023-3\_FundamentoEstadisticos/tarea\_01")

## Lee los datos

tips <- read\_csv("tips.csv")

glimpse(tips)

## Recodificar nombres y niveles

propinas <- tips %>%

rename(cuenta\_total = total\_bill,

propina = tip, sexo = sex,

fumador = smoker,

dia = day, momento = time,

num\_personas = size) %>%

mutate(sexo = recode(sexo, Female = "Mujer", Male = "Hombre"),

fumador = recode(fumador, No = "No", Si = "Si"),

dia = recode(dia, Sun = "Dom", Sat = "Sab", Thur = "Jue", Fri = "Vie"),

momento = recode(momento, Dinner = "Cena", Lunch = "Comida")) %>%

select(-sexo) %>%

mutate(dia = fct\_relevel(dia, c("Jue", "Vie", "Sab", "Dom")))

propinas

## 1. Calcula percentiles de la variable propina

## junto con mínimo y máxmo

quantile(propinas$propina, probs = seq(0, 1, 0.05))

## 2. Haz una gráfica de cuantiles de la variable propina

propinas <- propinas %>%

mutate(orden\_propina = rank(cuenta\_total, ties.method = "first"),

f = orden\_propina / n() )

## aquí tu código

# ggplot

ggplot(propinas, aes(x = cuenta\_total)) +

geom\_histogram()

## 3. Haz un histograma de la variable propinas

## Ajusta distintos anchos de banda

# ggplot con binwidth=[0.1 0.5 5]

ggplot(propinas, aes(x = cuenta\_total)) +

geom\_histogram(binwidth=5)

A graph of a number of objects

Description automatically generated

Ilustración 1: histogram with bandwidth 5

A graph of a graph showing the number of numbers

Description automatically generated with medium confidence

Ilustración 2: Histogram with bandwidth 1

A bar code graph with numbers

Description automatically generated

Ilustración 3: Histogram with bandwidth 0.1

## 4. Haz una gráfica de cuenta total contra propina

## ggplot

ggplot(propinas) +

geom\_point(aes(x = cuenta\_total, y = propina, color=num\_personas))

A graph showing a number of dots

Description automatically generated

Ilustración 4: cuenta vs. propina

A graph with a line and a blue line

Description automatically generated

Ilustración 5: cuenta vs propina usando alpha

A graph with black dots and blue line

Description automatically generated

Ilustración 6: cuenta vs propina usando size

A graph with blue dots and a blue line

Description automatically generated

Ilustración 7: cuenta vs propina usando coloue

ggplot(propinas,aes(x = cuenta\_total, y = propina, color=num\_personas)) +

geom\_point() +

geom\_smooth(method = "glm")

## 5. Calcula propina en porcentaje de la cuenta total

propinas <- propinas %>%

mutate(pct\_propina = propina/cuenta\_total)

## calcula algunos cuantiles de propina en porcentaje

quantile(propinas$pct\_propina, probs = seq(0, 1, 0.05))

## 6. Haz un histograma de la propina en porcentaje. Prueba con

## distintos anchos de banda.

## binwidth = [default, 0.1, 0.01]

ggplot(propinas) +

geom\_histogram(aes(x=pct\_propina), binwidth = 0.01) +

facet\_wrap(~momento)

A graph of a number of objects

Description automatically generated with medium confidence

Ilustración 8: Histogram with bandwidth 0.01

A graph of a number of objects

Description automatically generated with medium confidence

Ilustración 9: Histogram with bandwidth 0.1

## 7. Describe la distribución de propina en pct. ¿Hay datos atípicos?

max(propinas$pct\_propina)

mean(propinas$pct\_propina)

median(propinas$pct\_propina)

sd(propinas$pct\_propina)

## La distribucion de los porcentages de propina parece ser una distribución normal.

## con un valor promedio de 16% y und mediano de 15.5% y una devacion estandard de 6.1%

atypical <- propinas$pct\_propina>0.3

propinas$pct\_propina[atypical == TRUE]

## hay dos datos atípicos: mas que 30% (32.6%, 71% 41.7%)

## Lo interesante aqui es que tambien el valor de la cuenta total es relativamente bajo y

## el numero de personas: numero de personas: 1 2 2, y la cuenta total:$3.07 $7.25 $9.60.

## tambien la hora, las tres son en la tarde: puede ser que solamente tomaron algo.

propinas$num\_personas[atypical == TRUE]

propinas$cuenta\_total[atypical == TRUE]

propinas$momento[atypical == TRUE]

##8. Filtra los casos con porcentaje de propina muy altos.

## ¿Qué tipos de cuentas son? ¿Son cuentas grandes o chicas? --> arriba en 7.)

filter(propinas, pct\_propina>0.3)

## cuenta\_total propina fumador dia momento num\_personas orden\_propina f pct\_propina

## <dbl> <dbl> <chr> <fct> <chr> <dbl> <int> <int> <dbl>

## 1 3.07 1 Yes Sab Cena 1 1 1 0.326

## 2 7.25 5.15 Yes Dom Cena 2 4 4 0.710

## 3 9.6 4 Yes Dom Cena 2 14 14 0.417

## 9. Haz una diagrama de caja y brazos para

## propina en dolares dependiendo del momento (comida o cena)

## ¿Cuál parece más grande? ¿Por qué? Haz otras gráficas si es necesario.

ggplot(propinas) +

geom\_boxplot(aes(x=momento, y= pct\_propina), outlier.colour = "red")

## parece mas grande el boxplot de la cena. Una razon puede ser que las propinas de la cena

## contiene los atípicos

## quitando las propinas mayor a 30% y vuelvo a usar ggplot con esos datos.

propinas\_filtered <-filter(propinas, pct\_propina<0.3)

ggplot(propinas\_filtered) +

geom\_boxplot(aes(x=momento, y= pct\_propina), outlier.colour = "red")

A graph showing the difference between two people

Description automatically generated

Ilustración 10: Boxplot propina, incluyendo atipicos

A graph showing a diagram

Description automatically generated with medium confidence

Ilustración 11:Boxplot propina, sin atipicos