TALLER ANALISIS EXPLORATORIO Y PROCESAMIENTO

PRESENTADO POR:

JORGE ALBERTO INFANTE AVENDAÑO

JESUS DAVID SUAREZ PEÑA

PRESENTADO AL DOCENTE:

ALVARO AGUSTIN OÑATE BOWEN

UNIVERSIDAD POPULAR DEL CESAR

FACULTAD DE INGENIERÍA Y TECNOLOGÍAS

INGENIERIA DE SISTEMAS

VALLEDUPAR - CESAR

2020

1. Obtenga el conjunto German Credit del repositorio UCI Machine Learning Repository y cárguelo en Rstudio o Python.

- ✓ ***Obtención de datos:***

```
Estado_cuenta_corriente Duracion_mes Historial_credito Proposito Monto_credito C
A11 6 A34 A43 1169 A65 A75 4 A93 A101 4 A121 67 A143 A152 2 A173 1 A192 A201 1
A12 48 A32 A43 5951 A61 A73 2 A92 A101 2 A121 22 A143 A152 1 A173 1 A191 A201 2
A14 12 A34 A46 2096 A61 A74 2 A93 A101 3 A121 49 A143 A152 1 A172 2 A191 A201 1
A11 42 A32 A42 7882 A61 A74 2 A93 A103 4 A122 45 A143 A153 1 A173 2 A191 A201 1
A11 24 A33 A40 4870 A61 A73 3 A93 A101 4 A124 53 A143 A153 2 A173 2 A191 A201 2
A14 36 A32 A46 9055 A65 A73 2 A93 A101 4 A124 35 A143 A153 1 A172 2 A192 A201 1
A14 24 A32 A42 2835 A63 A75 3 A93 A101 4 A122 53 A143 A152 1 A173 1 A191 A201 1
A12 36 A32 A41 6948 A61 A73 2 A93 A101 2 A123 35 A143 A151 1 A174 1 A192 A201 1
A14 12 A32 A43 3059 A64 A74 2 A91 A101 4 A121 61 A143 A152 1 A172 1 A191 A201 1
A12 30 A34 A40 5234 A61 A71 4 A94 A101 2 A123 28 A143 A152 2 A174 1 A191 A201 2
A12 12 A32 A40 1295 A61 A72 3 A92 A101 1 A123 25 A143 A151 1 A173 1 A191 A201 2
A11 48 A32 A49 4308 A61 A72 3 A92 A101 4 A122 24 A143 A151 1 A173 1 A191 A201 2
A12 12 A32 A43 1567 A61 A73 1 A92 A101 1 A123 22 A143 A152 1 A173 1 A192 A201 1
A11 24 A34 A40 1199 A61 A75 4 A93 A101 4 A123 60 A143 A152 2 A172 1 A191 A201 2
A11 15 A32 A40 1403 A61 A73 2 A92 A101 4 A123 28 A143 A151 1 A173 1 A191 A201 1
A11 24 A32 A43 1282 A62 A73 4 A92 A101 2 A123 32 A143 A152 1 A172 1 A191 A201 2
A14 24 A34 A43 2424 A65 A75 4 A93 A101 4 A122 53 A143 A152 2 A173 1 A191 A201 1
A11 30 A30 A49 8072 A65 A72 2 A93 A101 3 A123 25 A141 A152 3 A173 1 A191 A201 1
A12 24 A32 A41 12579 A61 A75 4 A92 A101 2 A124 44 A143 A153 1 A174 1 A192 A201 2
A14 24 A32 A43 3430 A63 A75 3 A93 A101 2 A123 31 A143 A152 1 A173 2 A192 A201 1
A14 9 A34 A40 2134 A61 A73 4 A93 A101 4 A123 48 A143 A152 3 A173 1 A192 A201 1
A11 6 A32 A43 2647 A63 A73 2 A93 A101 3 A121 44 A143 A151 1 A173 2 A191 A201 1
A11 10 A34 A40 2241 A61 A72 1 A93 A101 3 A121 48 A143 A151 2 A172 2 A191 A202 1
A12 12 A34 A41 1804 A62 A72 3 A93 A101 4 A122 44 A143 A152 1 A173 1 A191 A201 1
```

- ✓ ***Carga de datos en Python:***

	Estado_cuenta_corriente	Duracion_mes	Historial_credito	Proposito	Monto_credito	Cuenta_ahorro_bonos	Empleo_actual	Tasa_pago_porcetaje	Estado_civil
0	A11	6	A34	A43	1169	A65	A75	4	
1	A12	48	A32	A43	5951	A61	A73	2	
2	A14	12	A34	A46	2096	A61	A74	2	
3	A11	42	A32	A42	7882	A61	A74	2	
4	A11	24	A33	A40	4870	A61	A73	3	

5 rows × 21 columns

2. Genere una hipótesis del origen y significado de los datos

- ✓ ***Numero de instancias:***

```
dataframe.count()
```

```
Estado_cuenta_corriente          1000
Duracion_mes                     1000
Historial_credito                1000
Proposito                        1000
Monto_credito                    1000
Cuenta_ahorro_bonos              1000
Empleo_actual                    1000
Tasa_pago_porcetaje              1000
Estado_civil_sexo                1000
Otros_deudores                   1000
Residencia                       1000
Propiedad                        1000
Edad                             1000
Planes_cuotas                    1000
Vivienda                         1000
Numero_creditos_banco            1000
Trabajo                          1000
Numero_personas_mantenimiento    1000
Telefono                         1000
Extranjero                       1000
Clasificacion                    1000
dtype: int64
```

Tiene un total de ***1000*** registros.

- ✓ ***Numero de atributos:***

```
dataframe.columns
```

```
Index(['Estado_cuenta_corriente', 'Duracion_mes', 'Historial_credito',
       'Proposito', 'Monto_credito', 'Cuenta_ahorro_bonos', 'Empleo_actual',
       'Tasa_pago_porcetaje', 'Estado_civil_sexo', 'Otros_deudores',
       'Residencia', 'Propiedad', 'Edad', 'Planes_cuotas', 'Vivienda',
       'Numero_creditos_banco', 'Trabajo', 'Numero_personas_mantenimiento',
       'Telefono', 'Extranjero', 'Clasificacion'],
      dtype='object')
```

```
len(dataframe.columns)
```

```
21
```

Tiene un total de 21 ***atributos***

- ✓ ***¿El conjunto de datos está etiquetado? ¿Cuantas clases tiene el conjunto de datos?***

Si, está etiquetado, tiene 1 clase que posee 2 etiquetas: 1 = bueno, 2=malo

- ✓ ***¿Cuantos atributos son numéricos y cuantos categóricos?***

```
dataframe.dtypes
```

```
Estado_cuenta_corriente          object
Duracion_mes                      int64
Historial_credito                object
Proposito                        object
Monto_credito                     int64
Cuenta_ahorro_bonos              object
Empleo_actual                    object
Tasa_pago_porcetaje               int64
Estado_civil_sexo                object
Otros_deudores                   object
Residencia                        int64
Propiedad                        object
Edad                              int64
Planes_cuotas                    object
Vivienda                         object
Numero_creditos_banco             int64
Trabajo                          object
Numero_personas_mantenimiento     int64
Telefono                         object
Extranjero                       object
Clasificacion                     int64
dtype: object
```

Contiene ***8 atributos numéricos*** y ***13 atributos categóricos*** para un total de 21 atributos.

- ✓ ***Reporte la moda para cada atributo categórico***

NOMBRE	MODA	SIGNIFICADO
Estado_cuenta_corriente	A14	No cuenta corriente
Historial_Credito	A32	créditos existentes pagados debidamente hasta ahora
Propósito	A43	Radio/televisión
Cuenta_ahorro_bonos	A61	…<100 DM
Empleo_actual	A73	1 <= …. <4
Estado_civil_sexo	A93	Soltero
Otros_deudores	A101	ninguno
Propiedad	A123	Si no A121 / A122: automóvil u otro, no en el atributo 6
Planes_cuotas	A143	ninguno

Vivienda	A152	Propio
Trabajo	A173	Empleado calificado /funcionario
Teléfono	A191	ninguno
Extranjero	A201	Si

- ✓ ***Reporte de la media, rango y desviación estándar para cada atributo***

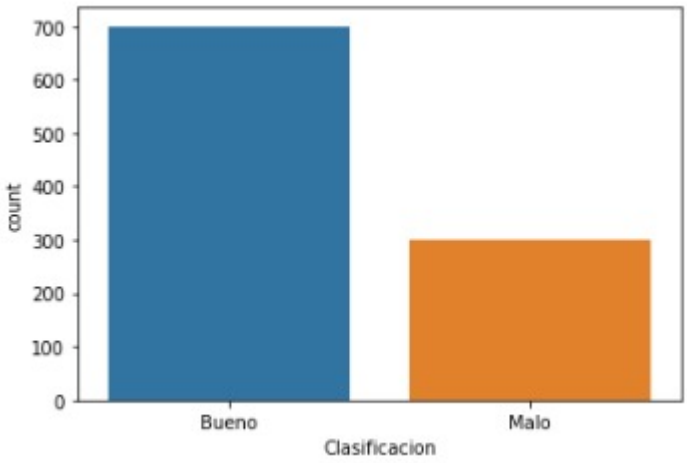
	Duracion_mes	Monto_credito	Tasa_pago_porcetaje	Residencia	Edad	Numero_creditos_banco	Numero_personas_mantenimiento	Clasificacion
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	20.903000	3271.258000	2.973000	2.845000	35.546000	1.407000	1.155000	1.300000
std	12.058814	2822.736876	1.118715	1.103718	11.375469	0.577654	0.362086	0.458487
min	4.000000	250.000000	1.000000	1.000000	19.000000	1.000000	1.000000	1.000000
25%	12.000000	1365.500000	2.000000	2.000000	27.000000	1.000000	1.000000	1.000000
50%	18.000000	2319.500000	3.000000	3.000000	33.000000	1.000000	1.000000	1.000000
75%	24.000000	3972.250000	4.000000	4.000000	42.000000	2.000000	1.000000	2.000000
max	72.000000	18424.000000	4.000000	4.000000	75.000000	4.000000	2.000000	2.000000

- ✓ ***Determinar el número de valores faltantes para cada atributo***

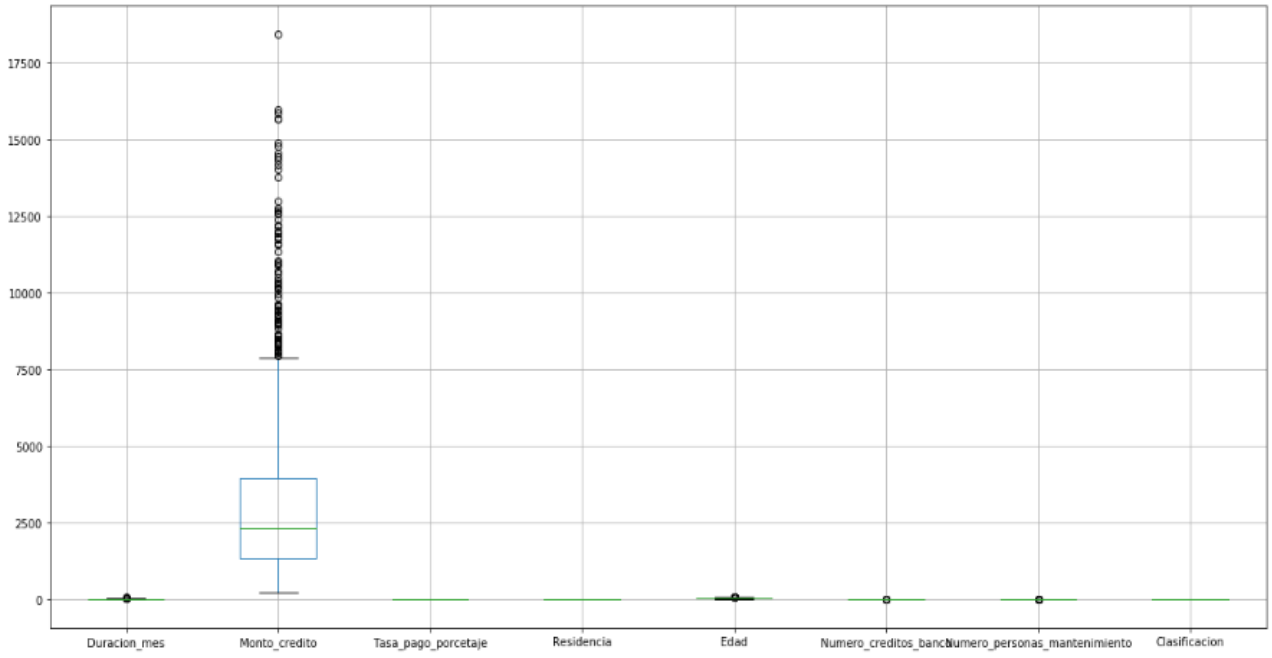
```
a
```

```
Estado_cuenta_corriente         0
Duracion_mes                    0
Historial_credito               0
Proposito                       0
Monto_credito                   0
Cuenta_ahorro_bonos             0
Empleo_actual                   0
Tasa_pago_porcetaje             0
Estado_civil_sexo               0
Otros_deudores                  0
Residencia                      0
Propiedad                       0
Edad                            0
Planes_cuotas                   0
Vivienda                        0
Numero_creditos_banco           0
Trabajo                         0
Numero_personas_mantenimiento   0
Telefono                        0
Extranjero                      0
Clasificacion                   0
dtype: int64
```

✓ ***Determine la distribución de las clases***

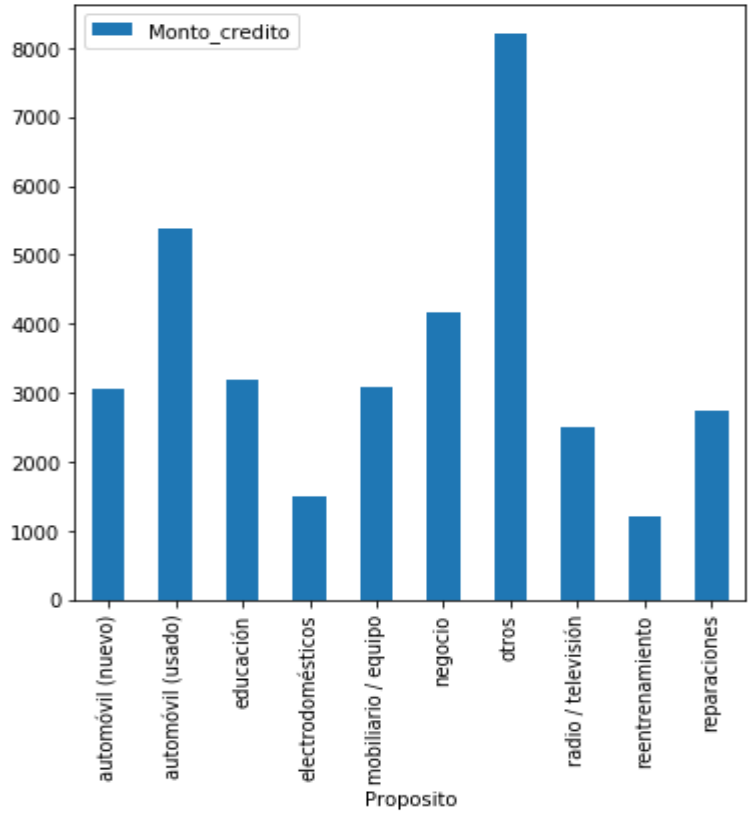


✓ ***Determine la existencia de datos atípicos***



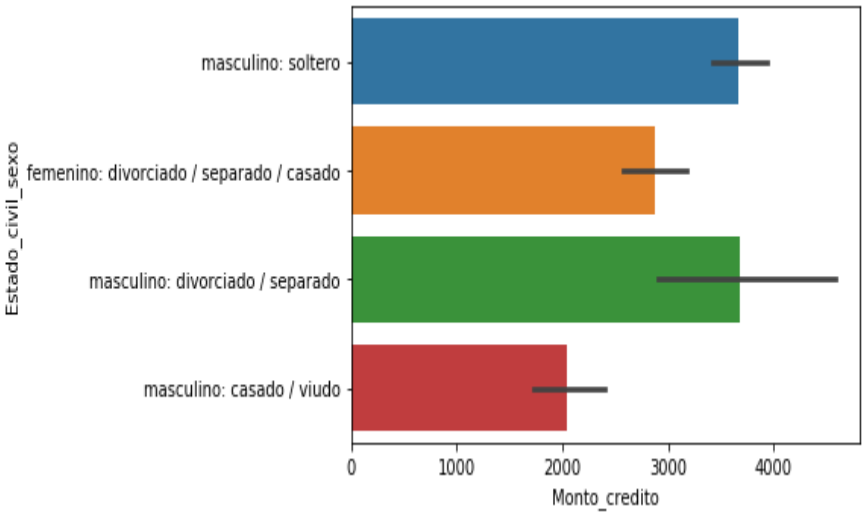
3. Responda las siguientes preguntas:

¿Cuál es lo propósito predominante de los préstamos?



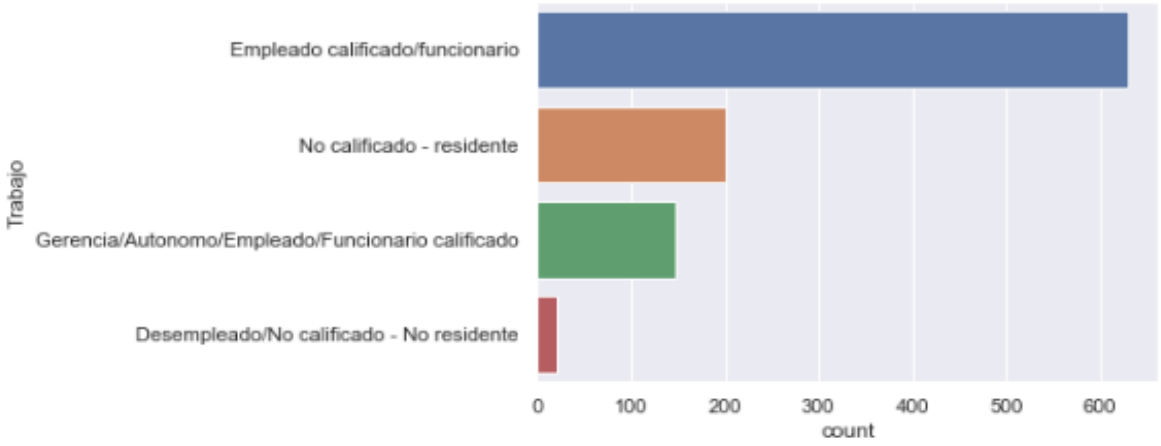
Las personas frecuentemente hacen préstamos para "otros" propósitos diferentes a los contemplados por el sistema.

¿Qué tipo de estatus tienen las personas que más hacen préstamos?



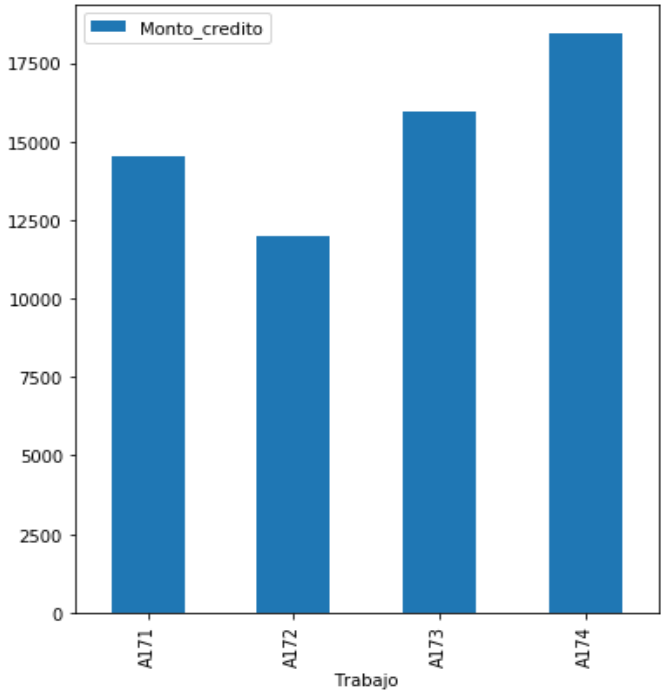
Masculino: soltero, es el estatus con mayor monto de créditos.

¿Y el perfil de la de menos préstamos?



Desempleado/No residente, es el perfil que menos número de préstamos hace.

¿Cuál es el perfil de las personas que hacen los prestamos más costoso?



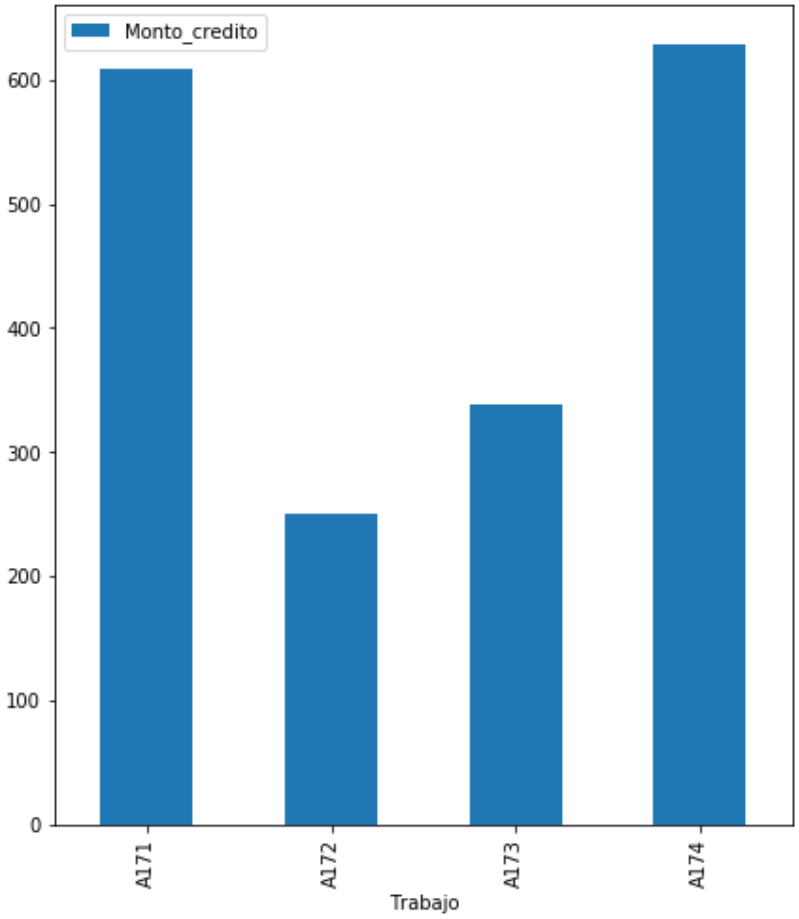
A171: desempleado / no calificado - no residente

A172: no calificado - residente

A173: empleado calificado / funcionario

A174: gerencia / autónomo /empleado / funcionario altamente calificado

¿Y el de los menos costosos?



A171: desempleado / no calificado - no residente

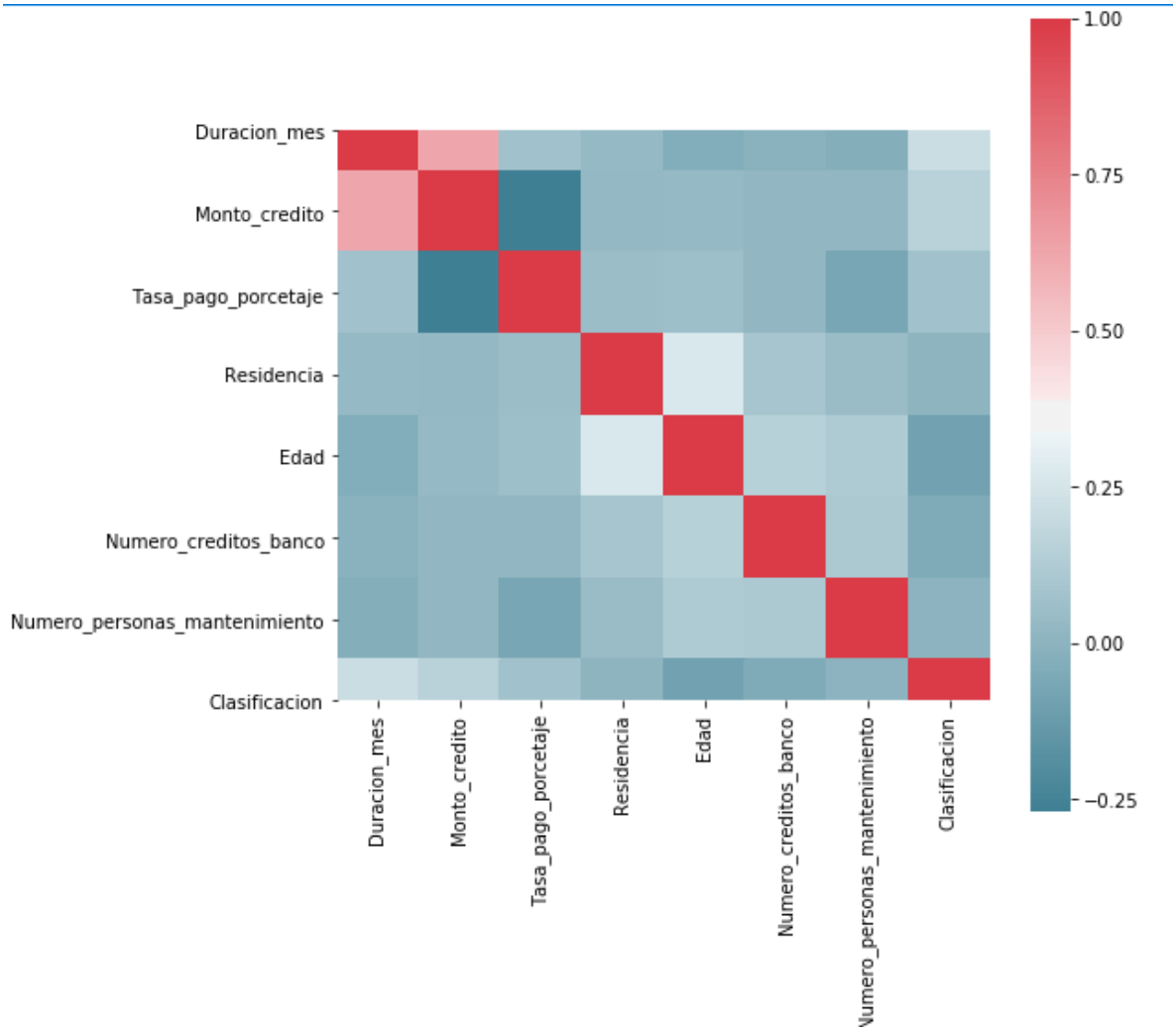
A172: no calificado - residente

A173: empleado calificado / funcionario

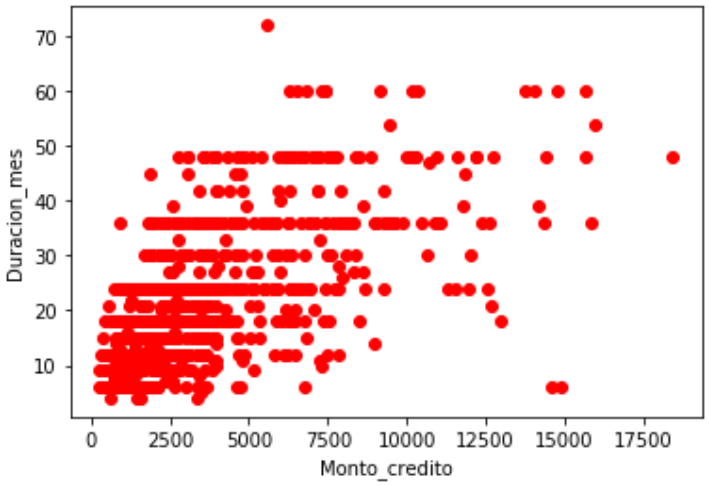
A174: gerencia / autónomo /empleado / funcionario altamente calificado

¿Puede establecer alguna relación entre edad, estatus personal y la clase?

¿Puede establecer alguna relación entre clase de trabajo, el número de créditos, estatus personal y la clase?



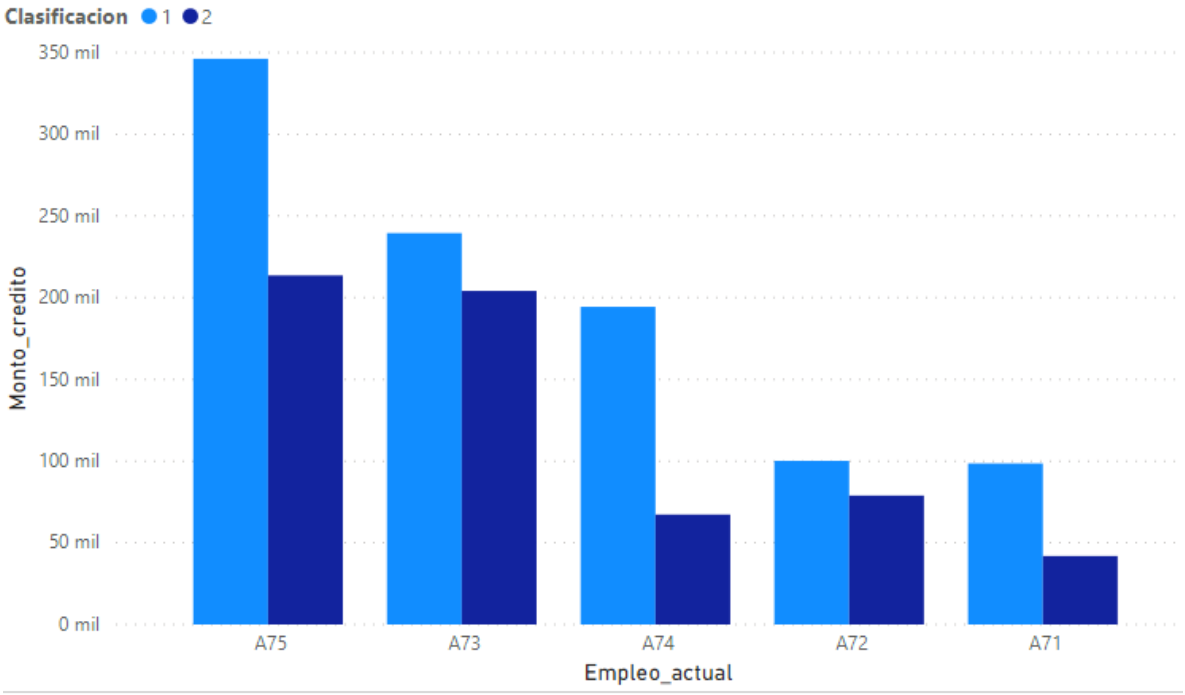
¿Existe alguna relación entre la cantidad solicitada y el número de meses del préstamo?



Entre las distintas relaciones esta es una de las mas fuertes ya que tiene mucho que ver la duración del préstamo con la cantidad y se observa que las personas obtienen la mayor cantidad de préstamos en una duración de mes baja ya que son prestamos a corto plazo

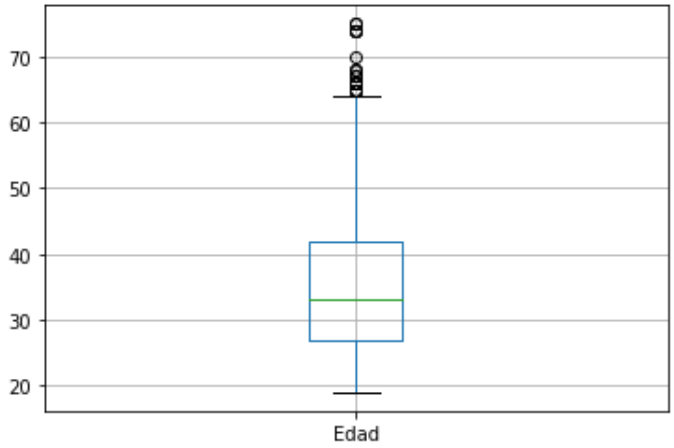
¿Existe alguna relación entre la edad, el estatus, la clase y la cantidad del préstamo?

7. Pruebe diferentes combinaciones entre los atributos y establezca las relaciones entre ellos, reporte la herramienta de visualización que utilizó para tal fin.



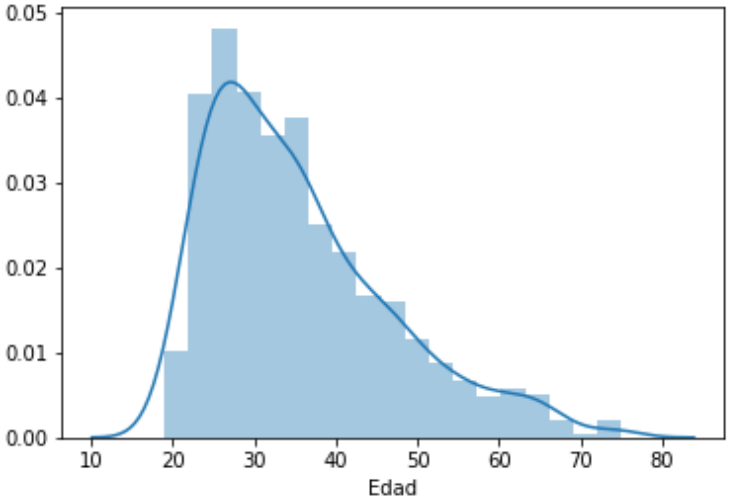
4. Realice los siguientes procedimientos sobre alguno de los atributos del conjunto de datos, analice los resultados y extraiga resultados

- Análisis de rangos intercuartiles

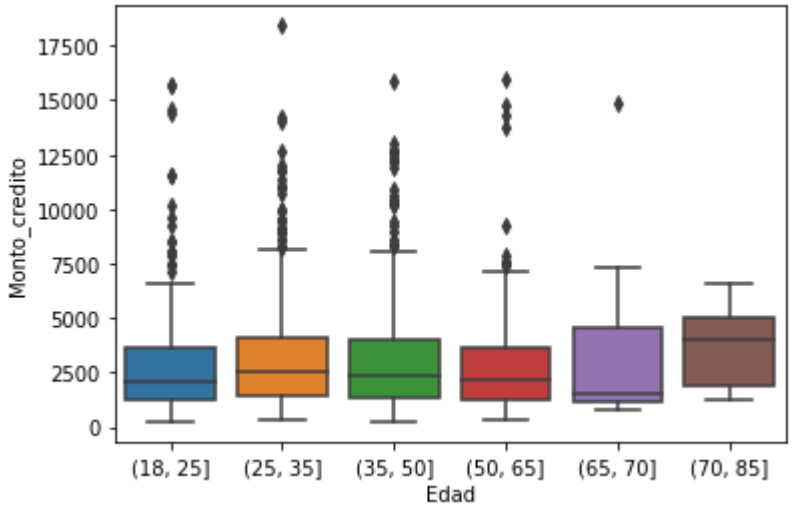


En esta grafica se observa los rangos de edad de las personas que adquieren un crédito donde la mediana es 35 años

- Histogramas

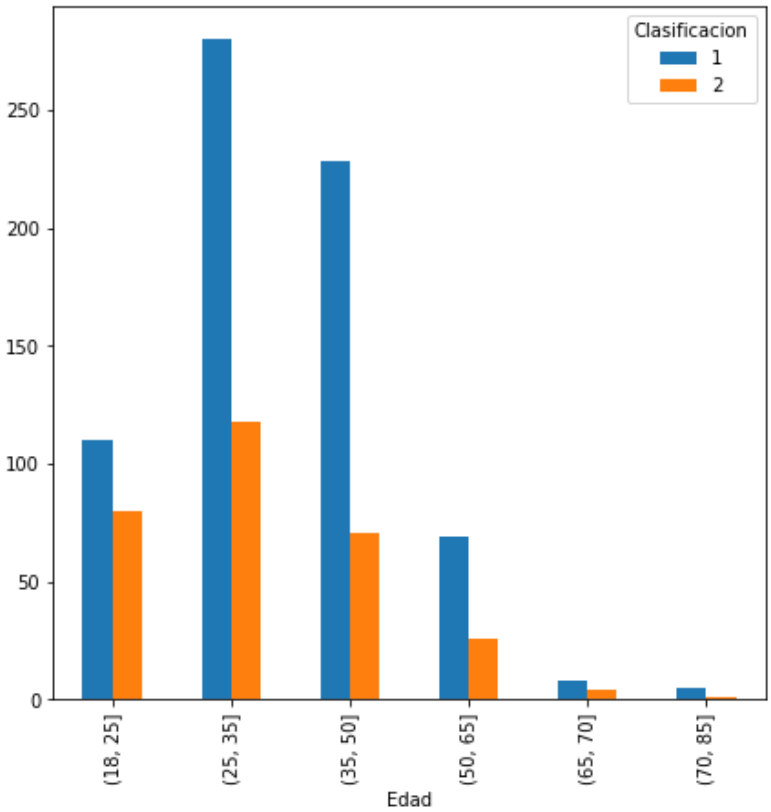


¿Qué grupo de personas según su edad hacen obtiene más créditos?



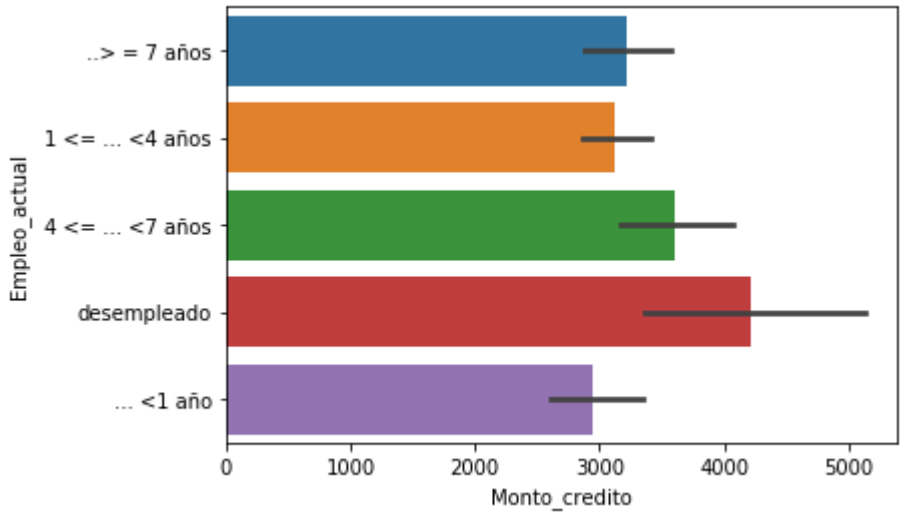
En la siguiente gráfica se puede observar que entre las personas que están entre los 18 y 50 años obtiene el monto de crédito más bajo, y ya que el promedio de edades está en ese rango re realizan más créditos, mientras que los grupos de edades mayor a 65 años hacen los prestamos más costosos

¿Qué rango de edades de los clientes tiene mejor clasificación?



En la siguiente gráfica se puede observar que las edades de los clientes más calificado están entre los 25 y 35 años ya que como la grafica anterior se ve que son las personas que hacen más créditos

¿Qué tiempo tienen los empleados actuales que hacen más prestamos?



En la gráfica se observa que el nivel más alto lo tienen las personas desempleadas y pues esto tiene una explicación ya que estas personas como en la vida cotidiana buscan como emprender un negocio o algo por el estilo