**ANÁLISIS ESTADÍSTICO DE UN DATASET BANCARIO**

**Problema:** Buscamos diferencias significativas entre los saldos de distintos tipos de clientes bancarios, con el fin que la unidad de marketing pueda realizar un análisis de segmentación de clientes.

Se proporciona un conjunto de muestras de clientes bancarios en un Dataset, que incluyen sus saldos bancarios y alguna información demográfica. La muestra de los clientes incluye a aquellos que tienen diferentes niveles de ingresos y por tanto pueden tener diferentes perfiles de gasto.

Se desea saber si existen diferencias significativas entre los saldos bancarios de los diferentes subsegmentos de su base de clientes, de modo que puedan diseñar productos específicos para los diferentes grupos (edad, trabajo, educación, estado civil, etc.).

**Contexto analítico:** El Dataset que utilizaremos en este caso fue recuperado de Kaggle (<https://www.kaggle.com/skverma875/bank-marketing-dataset>).

En este caso, daremos el primer paso para pasar del análisis exploratorio de datos al análisis confirmatorio.

El Dataset incluye el saldo de la cuenta de 45.211 clientes, junto con otras características relevantes que utilizaremos en este caso:

**balance**: saldo bancario; variable clave de interés

**job**: el título del trabajo (“management”, “technician”, etc)

**marital**: estado civil ("single", "married" or "divorced")

**education**: nivel de educación (("primary","secondary","tertiary","unknown")

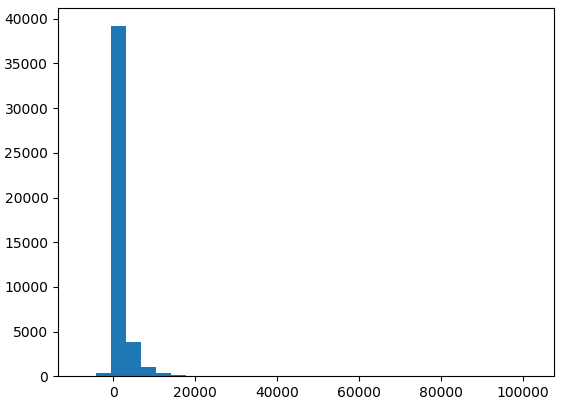
**default**: el cliente ha incumplido con sus pagos ("yes" o "no")

**loan**: el cliente ha obtenido un préstamo ("yes" o "no")

Nuestra primera búsqueda corresponderá en observar el rango de los saldos bancarios de todos los clientes

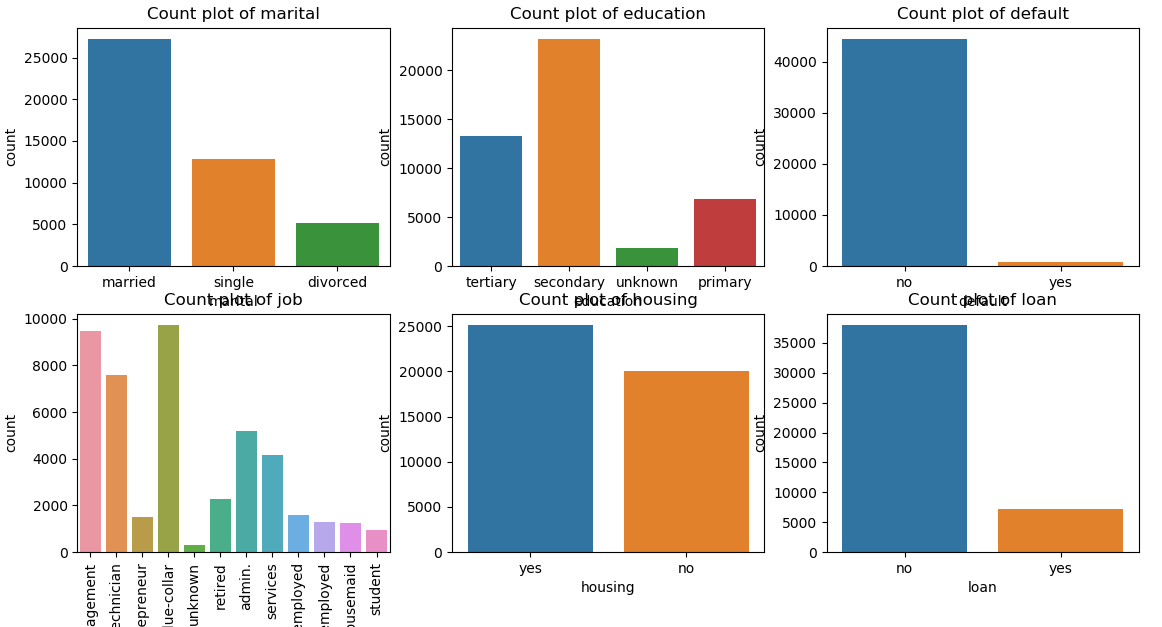
Del histograma figure\_1\_balance vemos un gran rango en el balance bancario de los clientes, que van desde pequeños saldos negativos, hasta saldos menores de $20.000.

Hay muchos subgrupos en la población de muestra. Por ejemplo, hay personas con diferentes grados de educación, perfiles de trabajo, etc. Por lo tanto, ahora observaremos los saldos bancarios de diferentes subgrupos.



figure\_1\_balance

Ahora, consideramos el número de clientes para las variables job, marital, education, default, housing, y loan.



Figure\_2\_client\_variables

Observamos que por categoria los clientes se distribuye de la siguiente manera:

Dentro de la categoria **marital** (estado civil), el subgrupo married (casado) representa la mayor cantidad de clientes, su número supera ampliamente a las subcategorías de single (soltero) y divorced (divorciado).

Dentro de la categoria **education** (educación), el subgrupo secondary (secundaria) representa la mayor cantidad de clientes, superando a las subcategorías de tertiary (estudios superiores) y primary (primaria). En esta categoria aparece un subgrupo unknown (desconocido), sin mas información al respeto, podemos sugerir que corresponde a errores en la captura de la información.

Dentro de la categoria **default** (incumplimiento de pagos), el subgrupo No representa la mayor cantidad de clientes, superando a las subcategorías de tertiary (estudios superiores) y primary (primaria). En esta categoria aparece un subgrupo unknown (desconocido), sin más información al respeto, podemos sugerir que corresponde a errores en la captura de la información.

Al encontrar el promedio del balance de las cuentas encontramos que = $ 1362.27, podemos compararlos con el promedio del periodo anterior, digamos, = $1341.12. Si la diferencia es positiva se podría pensar que el comportamiento del consumidor podría haber cambiado ligeramente y que, en promedio, los clientes están manteniendo más dinero en su cuenta bancaria.

Sin embargo, esta diferencia podría deberse a la variación estadística debido a la variabilidad del muestreo. Tendríamos que realizar una prueba más rigurosa para determinar si esta diferencia es realmente significativa después de tener en cuenta dicha variabilidad.

El procedimiento que utilizamos para ayudarnos a decidir si una diferencia entre la media de una población y un valor de referencia es estadísticamente significativa se denomina **prueba de hipótesis**. En nuestro contexto, es el saldo bancario promedio de los clientes este año y es su saldo bancario medio del año pasado.

El primer paso es definir la **hipótesis nula** (a menudo indicada como ). La hipótesis nula siempre corresponde a la hipótesis de no cambio, es decir, el estatus quo sigue siendo válido. Formalmente, esto se escribe como:

En nuestro caso, sería equivalente en la hipótesis de que el saldo promedio para toda la población de clientes es el mismo que el saldo promedio del año pasado, es decir = $1341.12. Debemos realizar una prueba estadística para evaluar la hipótesis de la gestión.

Las pruebas estadísticas reportan un valor . Esta es la cantidad clave que utilizaremos para determinar si el resultado de la prueba fue significativo. Introduzcamos la cantidad del nivel de significación .

Si esta probabilidad es menor que nuestro nivel de significación ( ) rechazamos y afirmamos que la diferencia observada es "estadísticamente significativa".

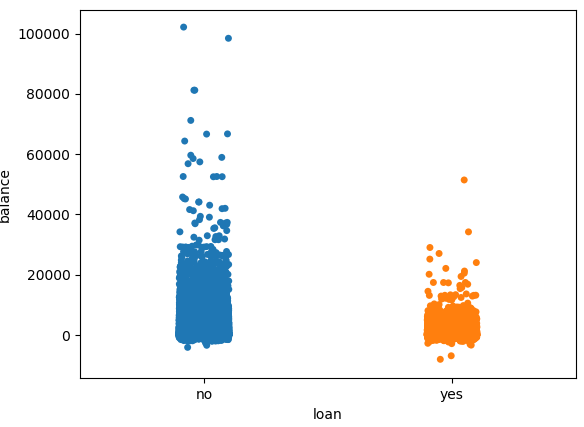
Si esta probabilidad es mayor que nuestro nivel de significación (𝑝>𝛼) tenemos que retener y afirmamos que la diferencia observada no es estadísticamente significativa.

En este caso, en este caso el valor encontrado es:

pvalue=0.13968331332845219 > 0,05

Concluimos que no hay suficiente evidencia estadística para decir con seguridad que es falso. En base a lo anterior no existe evidencia estadística para afirmar que los clientes han cambiado su comportamiento de ahorro con respecto al año anterior.

Ahora debemos determinar si existen diferencias entre los distintos grupos de clientes en cuanto a los saldos bancarios. Empezamos por examinar los saldos de dos subcategorías: los que pidieron un préstamo y los que no lo hicieron:



Figure\_3\_loan

Dado que la anteriorgrafica de distribución de los datos tiene colas pesadas, es difícil decir por medio del gráfico, si hay una diferencia entre las medias de los dos subgrupos. Continuamos nuestro análisis de datos observando las estadísticas de resumen para cada uno de los dos grupos:

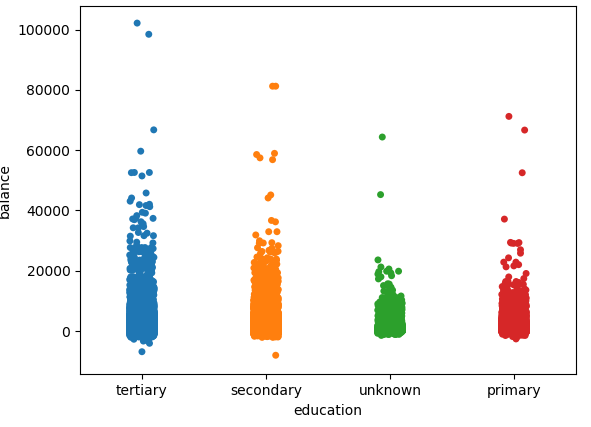
|  |  |
| --- | --- |
| Yes | No |
| count 7244.000000  mean 774.309912  std 1908.283253  min -8019.000000  25% 2.000000  50% 258.000000  75% 864.250000  max 51439.000000 | count 37967.000000  mean 1474.453631  std 3204.088951  min -4057.000000  25% 94.000000  50% 496.000000  75% 1558.000000  max 102127.000000 |

Podemos comprobar estadísticamente si las medias de los dos grupos son diferentes entre sí; es decir, si la diferencia entre el saldo medio de los grupos con un préstamo () es diferente del saldo medio del grupo sin préstamo ().

En esta caso el valor de 𝑝 - es muy pequeño (p = 2.76 e-1376); por lo tanto, rechazamos la hipótesis nula de que los dos grupos tengan la misma media. Es decir hay una diferencia estadisticamente significativa y por tanto los dos subgrupos son diferentes.

Igualemente, podemos identificar en este Dataset si la educación juega un papel en el comportamiento de los clientes. La variable educación tiene 4 categorías:

['tertiary' 'secondary' 'unknown' 'primary']

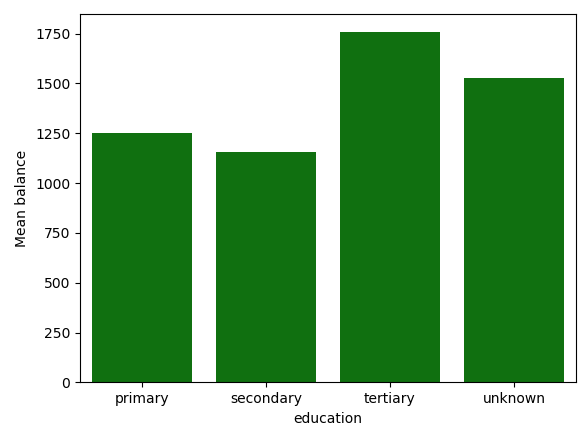


Figure\_4\_dispersion\_education

Visualizamos el balance por educación a través del gráfico dispersión anterior. La representación visual en este caso no proporciona demasiada información; por lo tanto, necesitamos un método alternativo para investigar más a fondo.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sum\_sq | df | F | PR(>F) |
| education | 3.220417e+09 | 3.0 | 116.682074 | 2.849538e-75 |
| Residual | 4.159034e+11 | 45207.0 | NaN | NaN |

Como podemos ver, al observar las estadísticas de 𝑝 - value que acompaña a 𝐹 – = (2.849538e-75) obtenemos un fuerte rechazo a la hipótesis nula, lo que nos lleva a concluir que los grupos educativos tienen diferencias en sus ingresos. El test ANOVA no nos dice qué existen pares de grupos que tienen medias diferentes entre sí. Para investigar estas diferencias más a fondo, vamos a reportar primero las estadísticas descriptivas por grupo, y luego mostrar las medias del grupo en un gráfico de barras:



Figure\_5\_education\_diferences

|  |  |
| --- | --- |
| descripci¾n de primary  count 6851.000000  mean 1250.949934  std 2690.743991  min -2604.000000  25% 61.000000  50% 403.000000  75% 1390.000000  max 71188.000000  Name: balance, dtype: float64 | descripci¾n de secondary  count 23202.000000  mean 1154.880786  std 2558.256739  min -8019.000000  25% 55.000000  50% 392.000000  75% 1234.000000  max 81204.000000  Name: balance, dtype: float64 |
| descripci¾n de tertiary  count 13301.000000  mean 1758.416435  std 3839.088305  min -6847.000000  25% 104.000000  50% 577.000000  75% 1804.000000  max 102127.000000  Name: balance, dtype: float64 | descripci¾n de unknown  count 1857.000000  mean 1526.754443  std 3152.228273  min -1445.000000  25% 106.000000  50% 568.000000  75% 1699.000000  max 64343.000000  Name: balance, dtype: float64 |

Cualitativamente parece haber una gran diferencia entre los niveles de educación terciaria y todos los niveles de educación inferior, pero la diferencia entre los grupos de educación primaria y secundaria es más difícil de evaluar. Queremos comprobar si hay una diferencia significativa en cada comparación por pares. Para probar esta secuencia de hipótesis podemos utilizar una variante del test 𝑡 –,el test por pares t- (pairwise t – test).

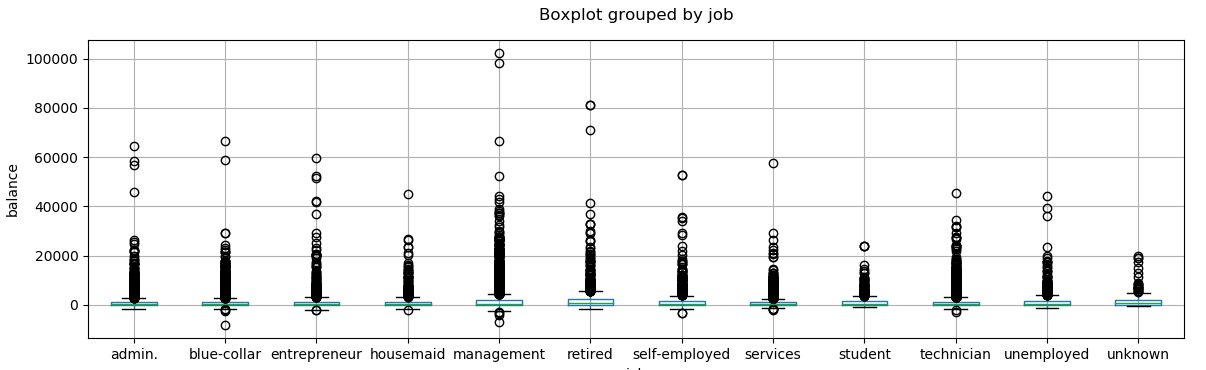
El resultado de la prueba es:

|  |  |  |
| --- | --- | --- |
|  | statistic | pvalue |
| tertiary vs. secondary | 16.18717929985782 | 1.4525474790099814e-58 |
| tertiary vs. unknown | 2.8825321831203117 | 0.003976238237673688 |
| tertiary vs. primary | 10.906634331433306 | 1.3023845080432282e-27 |
| secondary vs. unknown | -4.95482328681934 | 7.83111710943088e-07 |
| secondary vs. primary | -2.6255174227477864 | 0.0086639693661419 |
| unknown vs. primary | 3.445496944400963 | 0.0005789323657180873 |

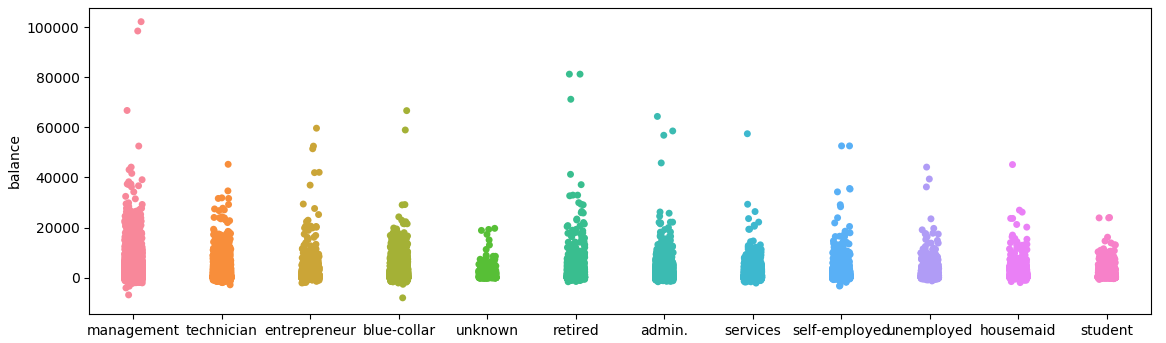
El resumen que se muestra en la tabla anterior indica que existe una diferencia significativa entre cada una de las comparaciones del par educación primaria y secundaria:

secondary vs. primary (p -- = 0.0086639693661419) < 0.05

Ahora vamos a realizar un análisis exploratorio de datos del saldo bancario por puesto de trabajo. Para ello primero recurrimos recurrimos a uan gradica de caja (boxplot) y a una grafica de dispersión.



Figure\_6\_grouped\_by\_job



Figure\_7\_grouped\_by\_job

Visualizamos el balance por trabajo a través de los gráficos anteriores. La representación visual en este caso no proporciona demasiada información; por lo tanto, utilizaremos nuevamente el test ANOVA nos dice qué par de grupos tienen medias diferentes entre sí..

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sum\_sq | df | F | PR(>F) |
| job | 4.341414e+09 | 11.0 | 43.007783 | 5.709430e-94 |
| Residual | 4.147824e+11 | 45199.0 | NaN | NaN |

Como podemos ver, al observar las estadísticas de 𝑝 - value que acompaña a 𝐹 – = (5.709430e-94) obtenemos un fuerte rechazo a la hipótesis nula, lo que nos lleva a concluir que los subgrupos de trabajo tienen diferencias en sus ingresos.

Una buena forma de comparar los valores de significancia de las pruebas de hipótesis nula por parejas para comparar las diferentes medias del grupo, aplicando la corrección de Bonferroni. Si embargo las librerias correspondiene no trabajan bajo un ambienete de pyhon 3.5.

**CONCLUSIONES**

Después de hacer un análisis exploratorio de los datos, introdujimos formalmente pruebas de hipótesis, podemos concluir los siguiente:

1. No hay diferencias significativas en el comportamiento de ahorro de los clientes.
2. Hay una diferencia estadisticamente significativa de saldo bancario en los dos subgrupos de quienes tienen prestamos y quienes no, favoreciendo ampliamentea a aquellos que no tienen prestamo.
3. Observamos que el nivel de educación afecta definitivamente el saldo bancario; los clientes con educación superior parecen tener una diferencia estadísticamente significativa en el saldo bancario en comparación con el resto de la población. Parece haber una diferencia estadísticamente significativa entre los clientes con educación primaria y secundaria.
4. También vimos bastantes diferencias estadísticamente significativas en los saldos bancarios de clientes con diferentes perfiles de trabajo.Aunque debido a limitaciones tecnicas no pudimos confirmar estas mediante librerias estadisticas como [pingouin-stats](https://pingouin-stats.org/).