UNIVERSIDAD ORT Uruguay

Facultad de Administración y Ciencias Sociales

Obligatorio

Principios de Estadística

Docente: Ec. Gastón Huertas

Integrante: Machado Cecilia – N° 213640

Fecha de entrega: 2 de diciembre de 2019

Índice

1 – Descripción de la base de datos

La base de datos utilizada en el siguiente informe fue obtenida del paquete ISLR y se llama *Credit*. La misma contiene información de balances de crédito de las tarjetas de 400 clientes. El objetivo del set de datos es predecir que clientes no van a pagar su deuda.
La base de datos contiene 400 elementos y tiene 11 variables que son:

- *Income*: ingreso en decenas de miles de dólares.
- *Limit*: límite de crédito.
- *Rating*: puntaje de crédito.
- *Cards*: número de tarjetas de crédito.
- *Age*: edad en años.
- *Education*: número de años de educación.
- *Gender*: Masculino o Femenino.
- *Student*: Si o No dependiendo si la persona fue estudiante.
- *Married*: Si o No dependiendo si la persona estuvo casada.
- *Ethnicity*: indicador de la etnia de la persona (Afro-Americano, Caucásico o Asiático).
- *Balance*: Promedio del balance de la tarjeta de crédito en dólares.

2 – Clasificación de las variables

Variables Cualitativas

Nominales	**Ordinales**
Género	
Estudiante	
Casado	
Etnia	

Variables Cuantitativas

Intervalo	**Razón**
	Ingresos
	Límite
	Tarjetas
	Edad
	Educación
	Saldo
	Rating

Todas las variables son discretas.

3 – Tipo de datos contenidos en la base

De corte transversal.

4 – Análisis Descriptivo

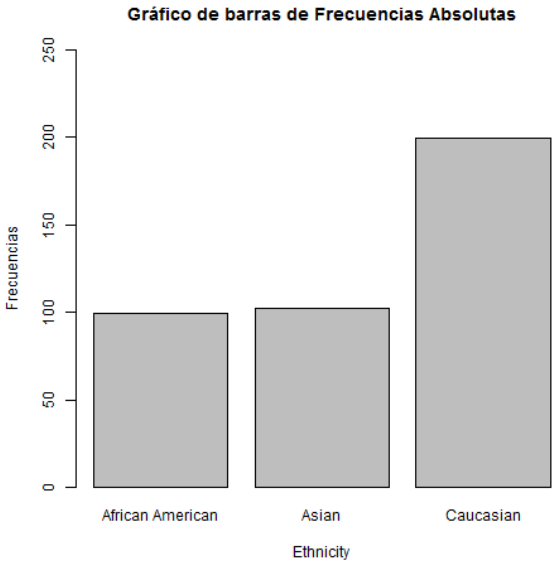
Variables cuantitativas elegidas: *Rating* e *Income.*
Variable cualitativa elegida: *Ethnicity.*

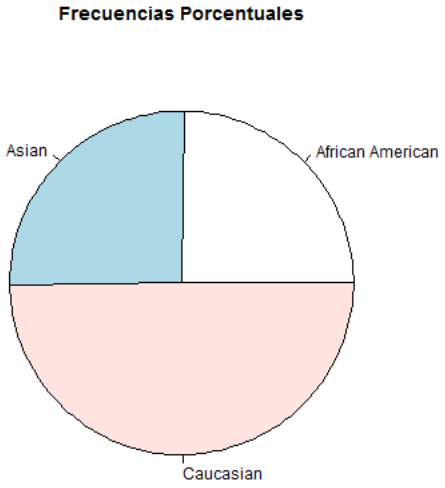
Los departamentos de riesgo de las empresas de crédito asignan un *Rating* que depende de varios factores. En el siguiente trabajo se busca analizar si hay una correlación entre el ingreso (*Income),* el Rating crediticio (*Rating*) y la etnia de la persona (*Ethnicity*). Considerando que Estados Unidos tiene una larga tradición multicultural, el objetivo es entonces determinar si hay prejuicios a la hora de asignar un determinado rating a un individuo.

a - Análisis de la variable cualitativa

Tabla de Frecuencias:

Ethnicity	FrecAbs	RelCredit	PorCredit	Frec_Abs_Acum	Frec_Rel_Acum	Frec_Por_Acum
African American	99	0.2475	24.75	99	0.2475	24.75
Asian	102	0.2550	25.50	201	0.5025	50.25
Caucasian	199	0.4975	49.75	400	10.000	100.00





Mediante la apreciación de la tabla de frecuencias y de los gráficos obtenidos podemos constatar que la etnia predominante dentro de la base de datos es *Caucasian*, presente en el 50% de las observaciones. Las restantes dos etnias, *African American y Asian*, representan el 50% restante de los datos pero están distribuidas en un valor cercano al 25% cada una.

b - Análisis de las variables cuantitativas

Variable Income

Tabla de frecuencias

clases_Income	Freq	Frec_rel	Frec_por	Frec_abs_acum	Frec_rel_acum	Frec_por_acum
(10,45.5]	266	0.665	66.5	266	0.665	66.5
(45.5,81]	82	0.205	20.5	348	0.870	87.0
(81,116]	28	0.070	7.0	376	0.940	94.0
(116,152]	16	0.040	4.0	392	0.980	98.0
(152,188]	8	0.020	2.0	400	1.000	100.0

El cálculo del ancho de clases se obtuvo de la siguiente manera:
Según teórico:

$$\text{Ancho de clases} = \frac{\text{Valor mayor} - \text{Valor menor}}{\text{Número de clases}}$$

Valor mayor: 186.6
Valor menor: 10.35
Lo recomendado es utilizar entre 5 y 20 clases, en éste caso utilizaremos 5.

$$\text{Ancho de clases} = \frac{186.6 - 10.35}{5}$$

⇩

$$\text{Ancho de clases} = 35.25$$

Se redondea el resultado obtenido, de lo contrario, algunas observaciones quedarían por fuera de las clases.

Ancho de clases = 35.50

Por lo tanto, la distribución de frecuencias de *Income* es:

- (10-45.5]
- (45.5-81]
- (81-116.5]
- (116.5-152]
- (152-188]

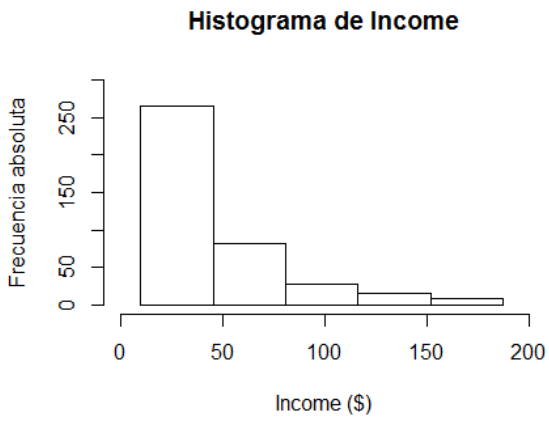
Medidas de tendencia

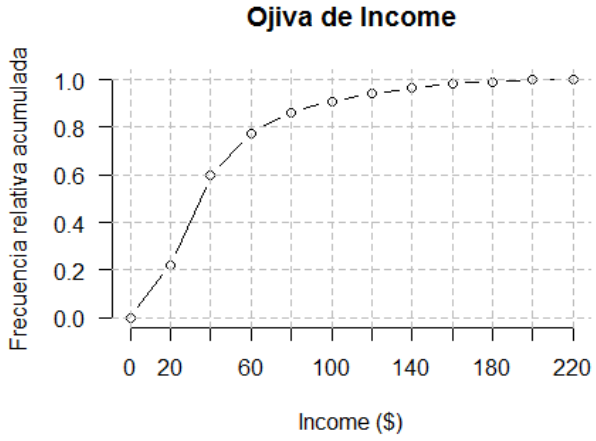
Tendencia	
Media	45.22
Mediana	33.12
Moda	23.79

Dispersión									
Cuartiles	21.01	33.12	57.47						
Deciles	14.58	19.21	23.76	27.81	33.12	40.08	51.96	63.83	92.45

Separación	
Rango	176.3
Rango interc.	36.46
Varianza	1242
Desviación estándar	35.24
Coeficiente de variación	77,94%

Gráficos:





Presencia o ausencia de observaciones atípicas:

Existen observaciones atípicas por encima del límite superior (112.2), son 29, y sus valores son los siguientes:

113.7	113.8	113.8	115.1	115.5	121.7	121.8	123.3	124.3	125.5
128.0	128.7	130.2	134.2	135.1	140.7	146.2	148.1	148.9	149.3
151.9	152.3	158.9	160.2	163.3	180.4	180.7	182.7	186.6	

No se encontraron observaciones atípicas por debajo del límite inferior (-33.69) debido a que el valor mínimo es 10.35.

Análisis general de la variable Income

Los valores máximos y mínimos de la variable son 186.6 y 10.35 respectivamente.
Los ingresos de los clientes que tiene mayor frecuencia se encuentran en el intervalo de entre 10 y 45.5 mil dólares, esto representa un 66,5% de los datos observados.

En el histograma se observa que la variable tiene una distribución asimétrica muy sesgada hacia la derecha, esto confirma que el valor de la media (45.22) se encuentre a la derecha de la mediana (33.12) y ésta última a la derecha de la moda (23.29).

La mayor diferencia de ingresos entre los clientes es de 176.3 mil dólares, asimismo, el 50% de los datos centrales tienen una diferencia máxima de precio que es de 36.46 mil dólares. En cuanto a la dispersión de los datos, podemos afirmar que es alta (77,9%), esto se debe a la presencia de cantidades importantes de valores atípicos, que a su vez, también afectaron el valor de la media y la desviación estándar.

Variable *Rating*

Tabla de frecuencias

clases_Rating	Freq	Frec_rel_Rating	Frec_por_Rating	Frec_abs_acum_Rating	Frec_rel_acum_Rating	Frec_por_acum_Rating
(90,270]	130	0.3250	32.50	130	0.3250	32.50
(270,450]	179	0.4475	44.75	309	0.7725	77.25
(450,630]	68	0.1700	17.00	377	0.9425	94.25
(630,810]	18	0.0450	4.50	395	0.9875	98.75
(810,990]	5	0.0125	1.25	400	10.000	100.00

El cálculo del ancho de clases se obtuvo de la siguiente manera:
Según teórico:

$$\text{Ancho de clases} = \frac{\text{Valor mayor} - \text{Valor menor}}{\text{Número de clases}}$$

Valor mayor:
Valor menor:
Lo recomendado es utilizar entre 5 y 20 clases, en éste caso utilizaremos 5.

$$\text{Ancho de clases} = \frac{982 - 93}{5}$$

$$\text{Ancho de clases} = 177{,}8$$

Se redondea el resultado obtenido, de lo contrario, algunas observaciones quedarían por fuera de las clases.

Ancho de clases = 180

Por lo tanto, la distribución de frecuencias de *Rating* es:

- (90-270]
- (270-450]
- (450-630]
- (630-810]
- (810-990]

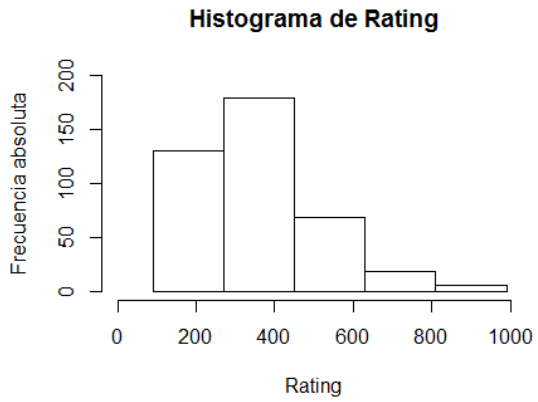
Medidas de tendencia

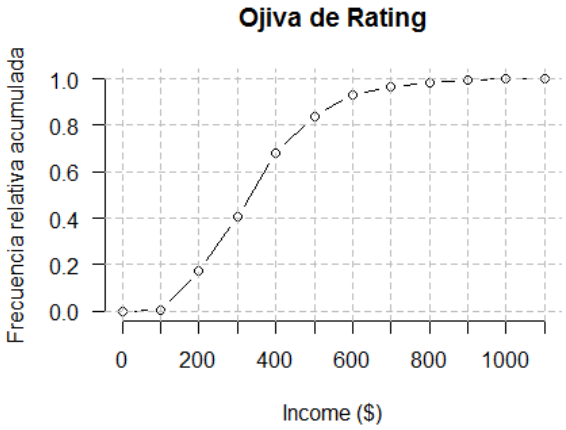
Tendencia	
Media	354.9
Mediana	344
Moda	344

Dispersión									
Cuartiles	247.2	344.0	437.2						
Deciles	167.0	216.6	263.7	299.0	344.0	377.0	410.0	466.4	549.5

Separación	
Rango	889
Rango interc.	190
Varianza	23940
Desviación estándar	154.7
Coeficiente de variación	43,59%

Gráficos





Presencia o ausencia de observaciones atípicas

Existen observaciones atípicas por encima del límite superior (722.2), son 11, y sus valores son los siguientes:

728	730	747	750	754	805
817	828	832	949	982	

No se encontraron observaciones atípicas por debajo del límite inferior (-37.8) debido a que el valor mínimo es 93.

Análisis general de la variable Rating

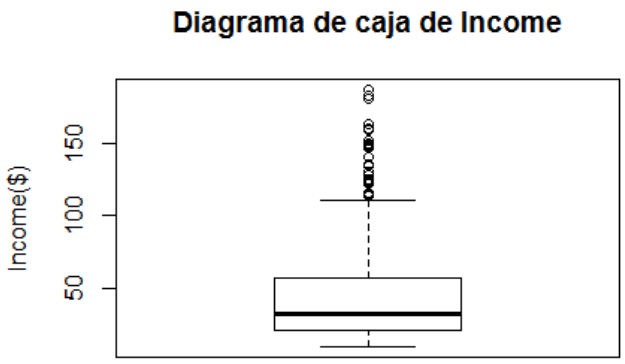
Los valores máximos y mínimos de la variable son 982 y 93 respectivamente.
El Rating de los clientes que tiene mayor frecuencia se encuentra en el intervalo de entre 270 y 450, esto representa un 44,75% de los datos observados. No obstante, hay una concentración de frecuencia importante en el intervalo de 90 a 270 representado por un 32,5% de los datos observados. El intervalo que posee menos representación es el de 810 y 990, teniendo a 5 clientes con ese puntaje de los 400 de la muestra, esto representa un 1,25%.

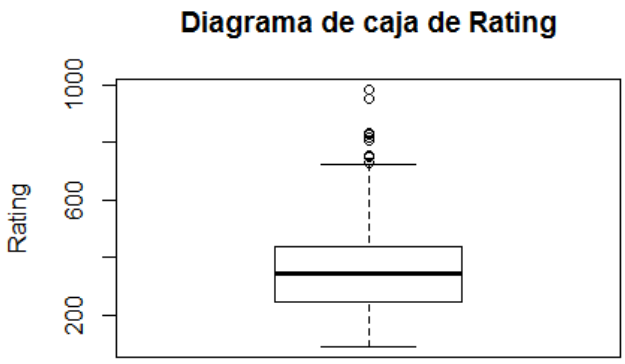
En el histograma se observa que la variable tiene una leve distribución asimétrica con sesgo a la derecha, esto confirma que el valor de la media (354.9) se encuentre a la derecha de la mediana (344) y ésta última con el mismo valor de la moda (344).

La mayor diferencia de Rating entre los clientes es de 889 puntos, asimismo, el 50% de los datos centrales tienen una diferencia máxima de precio que es de 190 puntos. En cuanto a la dispersión de los datos, podemos afirmar que es media (43,59%), esto se debe a la presencia de valores atípicos en la muestra.

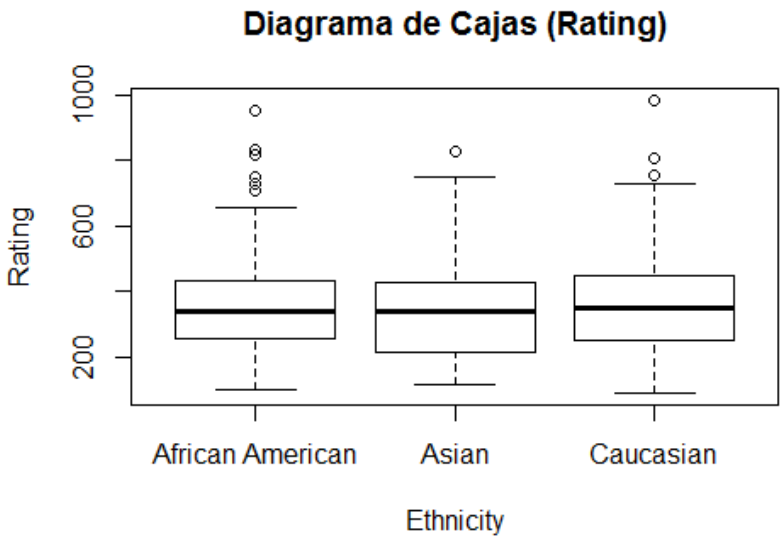
5 – Análisis de Distribución

Comparación de las dos variables cuantitativas

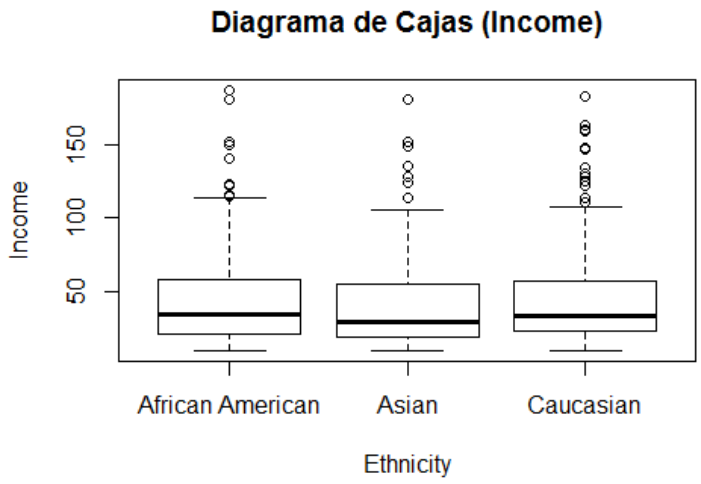




Distribución por grupos de cada variable cuantitativa por clases de la variable cualitativa.



- Se observa que el puntaje de *Rating* más alto corresponde a *Caucasian* y el más bajo a *Asian.*
- En base a las medianas, los valores de African American y Asian son muy similares y la mediana de Caucasian se despega levemente hacia la derecha.
- Existen observaciones atípicas de *Rating* en los 3 grupos, aunque *African American* es el que posee mayor cantidad de valores atípicos, y *Asian* tan solo un valor atípico.
- El Rating de *African American* parece tener menor variación, mientras que *Asian* parece tener mayor variación de puntaje.
- Las distribuciones de los puntajes de *African American* y *Caucasian* parecen bastante simétricas. La distribución de *Asian* parece asimétrica con rama a la izquierda.



- Se observa que el ingreso más alto corresponde a *African American* y el más bajo a *Asian.*
- En base a las medianas, los 3 grupos poseen valores similares entre sí, aunque la mediana de *Asian* tiene un valor levemente menor, esto es explicado por el percibimiento de menor ingreso como se mencionó en el punto anterior.
- Existen observaciones atípicas de *Income* en los 3 grupos, aunque *Caucasian* es el que posee mayor cantidad de valores atípicos, y *African American* el que posee menores cantidades de dichos valores.
- El ingreso de *Caucasian* parece tener menor variación, mientras que *Asian* parecen tener mayor variación de ingreso.
- Las distribuciones del ingreso de todos los grupos tienen asimetría con rama a la derecha, aunque las tendencias de las distribuciones de *Asian y Caucasian* parecen tener una asimetría más prominente.

6 – Análisis de Asociación

Tabulación cruzada entre Ethnicity *e* Income

Ethnicity/Income	(10,45.5]	(45.5,81]	(81,116]	(116,152]	(152,188]	Sum
African American	66	16	10	5	2	99
Asian	69	20	7	4	2	102
Caucasian	131	46	11	7	4	199
Sum	266	82	28	16	8	400

- Entre los 400 clientes, 199 de ellos son *Caucasian* que perciben ingresos, mientras que *African American* y *Asian* perciben 99 y 102 respectivamente. Por lo cual, la etnia que percibe más ingresos de la muestra son los *Caucasian* y la que percibe menores ingresos son los *African American.*
- Entre los 400 clientes, 131 de ellos son *Caucasian* que perciben entre 10 y 45.5 mil dólares, es decir, los ingresos más bajos y la etnia con mayor frecuencia en dicho intervalo.
- Entre los 400 clientes, 4 de ellos son *Caucasian* que perciben entre 152 y 188 mil dólares, es decir, los ingresos más altos y la etnia con mayor frecuencia en dicho intervalo.

Tabulación cruzada entre Ethnicity *e* Income *por distribución porcentual* [1]

Ethnicity/Income	(10,45.5]	(45.5,81]	(81,116]	(116,152]	(152,188]	Sum
African American	67	16	10	5	2	100
Asian	68	20	7	4	2	100
Caucasian	66	23	6	4	2	100
Sum	66	20	7	4	2	100

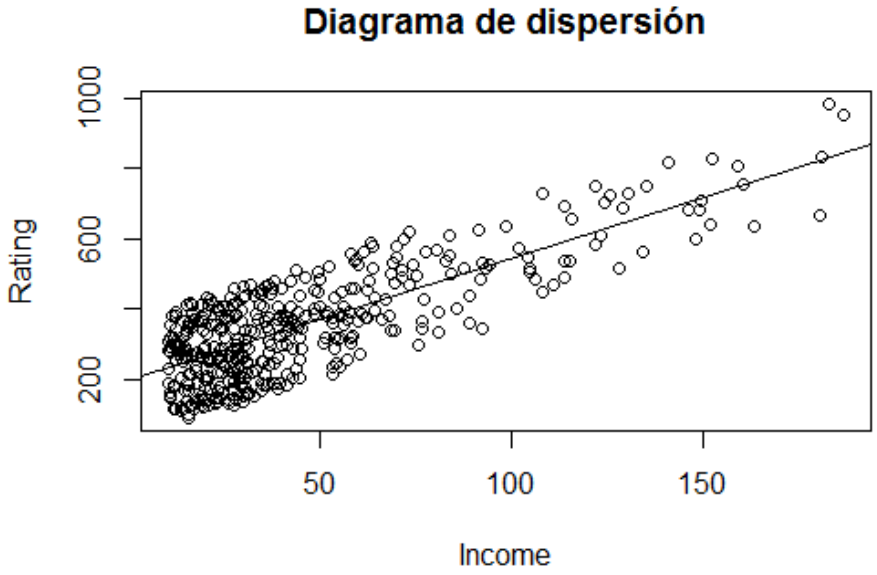
- Entre los clientes con ingresos más bajos, entre 10 y 45.5 mil dólares, los porcentajes mayores de concentración son para *Asian* con un 68%, aunque no hay demasiada diferencia con las restantes etnias ya que para *African American* y para *Caucasian* los porcentajes son de 67 y 66% respectivamente.
- Entre los clientes con los ingresos más altos, entre 152 y 188 mil dólares, los porcentajes de concentración es el mismo para las tres etnias, 2% para cada una de ellas.

Medidas de Asociación

Podemos afirmar que hay una asociación entre las variables *Ethnicity* e *Income,* es decir, que el valor de una de ellas ejerce influencia sobre la otra, debido a que el valor del Chi Cuadrado es mayor que cero (4.017). Pero la asociación entre ellas es débil ya que el valor del Índice de Cramer (0.07086) se encuentra entre los valores 0 y 0.3.

1 - Los datos que se visualizan son los porcentajes por filas.

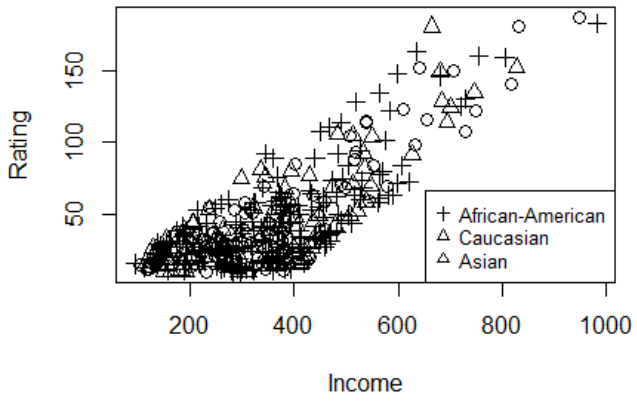
7 – Análisis de Correlación



Se observa que hay una tendencia a que cuanto más alto sea el ingreso de los clientes de la muestra, el rating crediticio de éstos también sea alto. Esto se puede comprobar al observar la pendiente positiva de la recta que está representada en diagrama de dispersión.
Para corroborar dicha correlación se calculó el valor de la Covarianza (4315), al ser positiva ésta sugiere que hay una relación lineal creciente. También se calculó el Coeficiente de Correlación Lineal de Pearson para evidenciar si dicha relación está perfectamente alineada o no con la recta creciente. El valor de dicho Coeficiente (0.7914) nos indica que la relación no está perfectamente alineada, pero si muestra que la correlación entre ambas variables cuantitativas es importante.

7.1 – BONUS

Gráfico de varios cuadrantes el diagrama de dispersión de las dos variables cuantitativas para cada una de las categorías que toma la variable cualitativa[2].



8 – Conclusión

Del análisis y tratamiento estadístico cabe destacar las siguientes observaciones y conclusiones:

El cliente con mayores ingresos pertenece a la etnia *Afro American,* sin embargo, no es el que tiene mayor *Rating* ni tampoco su etnia en general es la que posee mayor concentración de ingresos.
Quien está mayormente presente en el set de datos es la etnia *Caucasian*, con un 50% de representación de los datos de la muestra, es por esto que posee la mayor cantidad de clientes dentro del intervalo más bajo de ingresos, aunque también tiene a la mayor cantidad de clientes con los ingresos más altos. No obstante, si se observa al grupo aislado de *Asian* se denota que es éste quien tiene mayor concentración de personas que perciben menores ingresos.
En cuanto al *Rating* quienes tienen mejores puntuaciones crediticias son los *Caucasian,* pero se observan bastantes valores atípicos dentro de la etnia *Afro American.*
También se pudo constatar una asociación entre la etnia y los ingresos, aunque la relación entre ambas es débil. No es el caso de *Income* y *Rating*, que tienen una importante correlación entre ambas y la misma es lineal positiva.

Por lo tanto, se puede concluir que el rating crediticio se ve más afectado por el ingreso de un cliente y también en parte, por su etnia.

[2] No se graficó más de un cuadrante debido a la ausencia de datos negativos en el set de datos.