Instituto Politecnico Nacional
Escuela Superior de Computo
Trabajadores en Estados Unidos

Objetivo

El propósito principal de la investigación presente es limpiar el conjunto de datos, este contiene datos censales extraídos de las Encuestas de Población Actual de los años 1994 y 1995 realizadas por la Oficina del Censo de los Estados Unidos, para poder determinar, comprender e identificar el comportamiento de los trabajadores en diversas áreas geográficas, así como los factores que influyen en su vida laboral. Dentro de los campos que localizamos en nuestro conjunto de datos encontramos cuestiones como razones de desempleo, si eran o no inmigrantes, si aun estudiaban o cual había sido su ultimo grado escolar, cuantos de sus familiares aún son menores de 18 años, entre otras cuestiones interesantes que nos pueden mostrar la situación general en la que se encontraban los trabajadores.
Para cumplir nuestro objetivo pondremos en practica los conceptos, las estrategias de limpieza de datos y las herramientas vistas en la unidad de aprendizaje de Data Mining, también la aplicación de técnicas como imputación, normalización y transformación, siempre justificando su decisión, por qué se procedió con cierta técnica o por qué se optó por una herramienta. De la misma manera usaremos gráficas de los datos más representativos para poder ilustrar los resultados obtenidos y dar una mejor interpretación de estos.

Descripción Conjunto de Datos

El conjunto de datos consiste en 40 atributos, de los cuales 7 son continuos y 33 son nominales. Incluyen variables demográficas y relacionadas con el empleo, sus datos fueron extraídos de las Encuestas de Población Actual de los años 1994 y 1995 realizadas por la Oficina del Censo de los Estados Unidos.
Con el fin de realizar un estudio completo se utilizaran todos los registros del conjunto resultando en un total de 199,523 registros a tratar, pueden ser consultados en https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
En seguida describiremos la información del conjunto de datos:
Propietario Original:
U.S. Census Bureau
http://www.census.gov/
United States Department of Commerce
Donante:
Terran Lane and Ronny Kohavi
Data Mining and Visualization
Silicon Graphics.
terran '@' ecn.purdue.edu, ronnyk '@' sgi.com

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

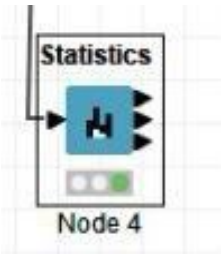
15 de Abril, 2021

Instituto Politecnico Nacional
Escuela Superior de Computo
Trabajadores en Estados Unidos

Selección de los atributos

Para hacer la seleccion de atributos que usaremos durante la investigacion nos basamos en dos factores importantes. El primer filtro fue la cantidad de valores faltantes y el segundo fue la relevancia que tienen para el estudio.
Para el primer filtro se empleo el nodo llamado Statics, en el podemos visualizar imformación importante de los datos, uno de estos datos importantes es la cantidad de valores faltantes. A continuación se mostrara un ejemplo de como se descartaron los datos con gran numero de datos faltantes:

Nodo Empleado



Informacion del atributo

Notemos que tenemos una cantidad muy grande de datos faltantes, por esta razon se descartó, una alta manipulación para completar estos datos ensuciarian los resultados del estudio a realizar.
De esta manera a continuación enlistaremos los 20 atributos que pasaron los dos filtros propuestos:

AAGE, ACLSWKR, AHGA, AHRSPAY,AHSCOL, AMARITL, AMJIND, ARACE, ASEX,AUNTYPE, CAPGAIN, CAPLOSS, NOEMP, PARENT, PEFNTVTY, PEMNTVTY, PENATVTY, PRCITSHP, WKSWORK, VETYN.

Con esta eleccion procederemos a desarrollar el diccionario de datos donde definiremos, de cada atributo seleecionado, su nombre, el tipo de dato, su dominio y una breve descripcion del mismo para un mejor entendimiento del estudio.

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena
Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional
Escuela Superior de Computo
Trabajadores en Estados Unidos

Diccionario de datos

Nombre del Atributo	AAGE
Tipo de dato	Númerico
Dominio	0-90
Descripción	Edad de la persona al momento del registro de la información
Nombre del Atributo	**ACLSWKR**
Tipo de dato	Nominal
Valores	Not in universe, Federal government, Local government, Never worked, Private, Self-employed-incorporated, Self-employed-not incorporated, State government, Without pay
Descripción	Hace referencia a la clase de trabajador que es la persona al momento del registro de la información
Nombre del Atributo	**AHGA**
Tipo de dato	Nominal
Valores	Children 7th and 8th grade 9th grade 10th grade High school graduate 11th grade 12th grade no diploma 5th or 6th grade Less than 1st grade Bachelors degree(BA AB BS) 1st 2nd 3rd or 4th grade Some college but no degree Masters degree(MA MS MEng MEd MSW MBA) Associates degree-occup /vocational Associates degree-academic program Doctorate degree(PhD EdD) rof school degree (MD DDS DVM LLB JD)
Descripción	Grado de estudio que tiene el individuo
Nombre del Atributo	**AHRSPAY**
Tipo de dato	Númerico
Dominio	0-9,999
Descripción	Haace referencia al salario por hora del individuo
Nombre del Atributo	**AHSCOL**
Tipo de dato	Nominal

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Nombre del Atributo	AHSCOL
Valores	Not in universe High school College or university
Descripción	Hace referencia a si el inidividuo estaba o no estudiando en al momento del registro de la información

Nombre del Atributo	AMARITL
Tipo de dato	Nominal
Valores	Never married Married-civilian spouse present Married-spouse absent Separated Divorced Widowed Married-A F spouse present
Descripción	Hace referencia al estado civil del individuo al momento del registro de la información

Nombre del Atributo	AMJIND
Tipo de dato	Nominal
Valores	Not in universe or children Entertainment Social services Agriculture, Education Public administration Manufacturing-durable goods Manufacturing-nondurable goods Wholesale trade, Retail trade Finance insurance and real estate Private household services Business and repair services Personal services except private HH Construction, Medical except hospital Other professional services, Transportation Utilities and sanitary services Mining Communications Hospital services Forestry and fisheries Armed Forces.
Descripción	Hace referencia al enfoque de la industria en la que se desenvuelve el individuo

Nombre del Atributo	ARACE
Tipo de dato	Nominal

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Nombre del Atributo	ARACE
Valores	White Black Other Amer Indian Aleut or Eskimo Asian or Pacific Islander
Descripción	Hace referencia a la raza del individuo
Nombre del Atributo	**ASEX**
Tipo de dato	Nominal
Valores	Female Male
Descripción	Hace referencia al genero biologico del individuo
Nombre del Atributo	**AUNTYPE**
Tipo de dato	Nominal
Valores	Not in universe Re-entrant Job loser - on layoff New entrant Job leaver Other job loser.
Descripción	Hace referencia a si el individuo tiene o no empleo
Nombre del Atributo	**CAPGAIN**
Tipo de dato	Numérico
Dominio	0 - 99,999
Descripción	Hace referencia a las ganancias de capital del individuo
Nombre del Atributo	**CAPLOSS**
Tipo de dato	Numérico
Dominio	0 - 4,608
Descripción	Hace referencia a las perdidas de capital del individuo
Nombre del Atributo	**NOEMP**
Tipo de dato	Numérico
Dominio	0-6
Descripción	Hace referencia a la cantidad de personas que trabajan para el empleador
Nombre del Atributo	**PARENT**
Tipo de dato	Nominal
Valores	Both parents present Neither parent present Mother only present Father only present Not in universe.
Descripción	Hace referencia a los miembros de la familia que son menores a 18 años

---

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena
Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Nombre del Atributo	PEFNTVTY
Tipo de dato	Nominal
Valores	Mexico United-States Puerto-Rico Dominican-Republic Jamaica Cuba Portugal Nicaragua Peru Ecuador Guatemala Philippines Canada Columbia El-Salvador Japan England Trinadad&Tobago Honduras Germany Taiwan Outlying-U S (Guam USVI etc) India Vietnam China Hong Kong Cambodia France Laos Haiti South Korea Iran Greece Italy Poland Thailand Yugoslavia, Holand-Netherlands Ireland Scotland Hungary Panama
Descripción	Hace referencia al lugar de nacimiento del papá del individuo

Nombre del Atributo	PEMNTVTY
Tipo de dato	Nominal

Nombre del Atributo	PEMNTVTY
Valores	India Mexico United-States Puerto-Rico Dominican-Republic England Honduras Peru Guatemala Columbia El-Salvador Philippines France Ecuador Nicaragua Cuba Outlying-U S (Guam USVI etc) Jamaica South Korea China Germany Yugoslavia Canada Vietnam Japan Cambodia Ireland Laos Haiti Portugal Taiwan Holand-Netherlands Greece Italy Poland Thailand Trinadad&Tobago Hungary Panama Hong Kong Scotland Iran
Descripción	Hace referencia al lugar de nacimiento de la mamá del individuo

Nombre del Atributo	PENATVTY
Tipo de dato	Nominal

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena
Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Nombre del Atributo	PENATVTY
Valores	United-States Mexico Puerto-Rico Peru Canada South Korea India Japan Haiti El-Salvador Dominican-Republic Portugal Columbia England Thailand Cuba Laos Panama China Germany Vietnam Italy Honduras Outlying-U S (Guam USVI etc) Hungary Philippines Poland Ecuador Iran Guatemala Holand-Netherlands Taiwan Nicaragua France Jamaica Scotland Yugoslavia Hong Kong Trinadad&Tobago Greece Cambodia Ireland
Descripción	Hace referencia al lugar de nacimiento del individuo

Nombre del Atributo	PRCITSHP
Tipo de dato	Nominal

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Nombre del Atributo	PRCITSHP
Valores	Native- Born in the United States Foreign born- Not a citizen of U S Native- Born in Puerto Rico or U S Outlying Native- Born abroad of American Parent(s) Foreign born- U S citizen by naturalization.
Descripción	Hace referencia a la ciudadania del individuo
Nombre del Atributo	**WKSWORK**
Tipo de dato	Nominal
Dominio	0-53
Descripción	Hace referencia a la cantidad de semanas que trabaja el indiviuo al año
Nombre del Atributo	**YEAR**
Tipo de dato	Nominal
Valores	94, 95
Descripción	Como fue un estudio de dos años, hace referencia al año en el que se realizo

Datos descriptivos

En esta seccion enlistaremos el mínimo, máximo, media, varianza, desviación estándar, los cuartiles y/o centiles y la descripción e interpretación correspondientes para comprender la naturaleza de los datos tratados.

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

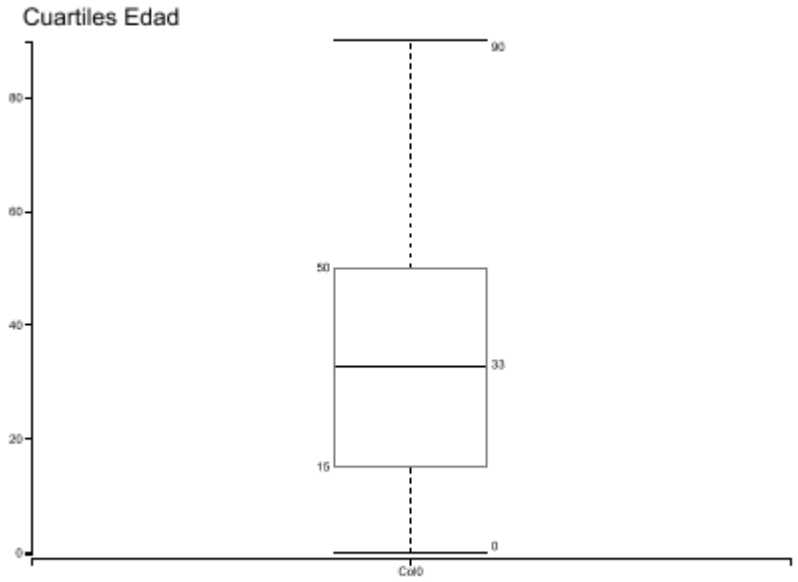
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

AAGE		
Dato	Valor	Interpretación
Mínimo	0	Se encuestaron incluso a bebés menores a un año.
Máximo	90	Los ciudadanos más grandes tenían 90 años.
Media	34.494	El promedio de edad de los encuestados fue 34 años.
Desviación estándar	22.311	
Varianza	497.776	
Cuartiles		



Descripción	Tanto en la tabla como en la imágen podemos ver que la edad minima que tuviero las personas que formaron parte de la encuesta tenian 0 años y la maxima edad fueron de 90. Enfocandonos en los cuartiles notemos que la mayor parte del volumen total se concentra en el intervalo de 15 a 50.

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

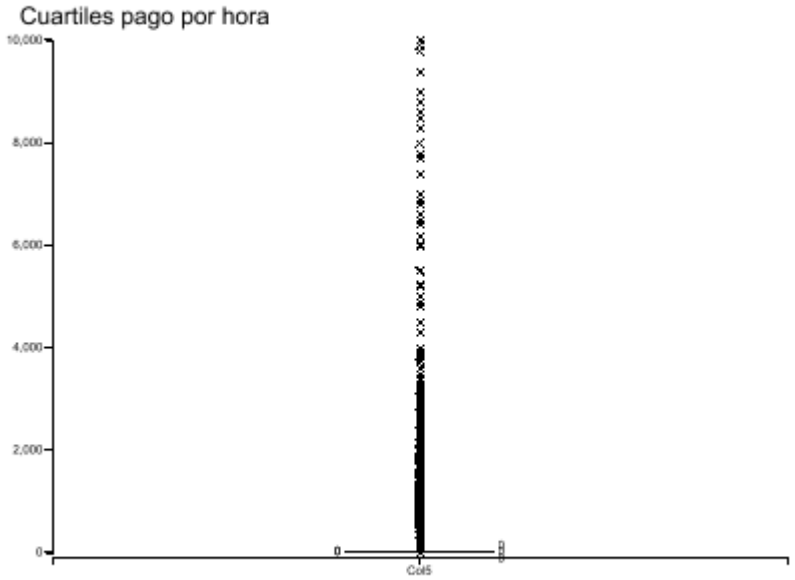
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

AHRSPAY		
Dato	Valor	Interpretación
Mínimo	0	Cantidad por hora mínima pagada.
Máximo	99,999	Cantidad por hora máxima pagada.
Media	55.427	Cantidad por hora promedio pagada.
Desviación estándar	274.896	
Varianza	75,568.06	
Cuartiles		



Descripción

Para la cantidad que se paga por hora tenemos que el minimo es de 0, correspondiendo por ejemplo a las personas del conjunto que no tiene un salario, siendo el maximo de 99,999,
Si obsevamos la grafica de caja notamos que el los salarios bajos se concentra la mayor frecuencia, los que llegan a los valores maximos son valores atipicos.

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

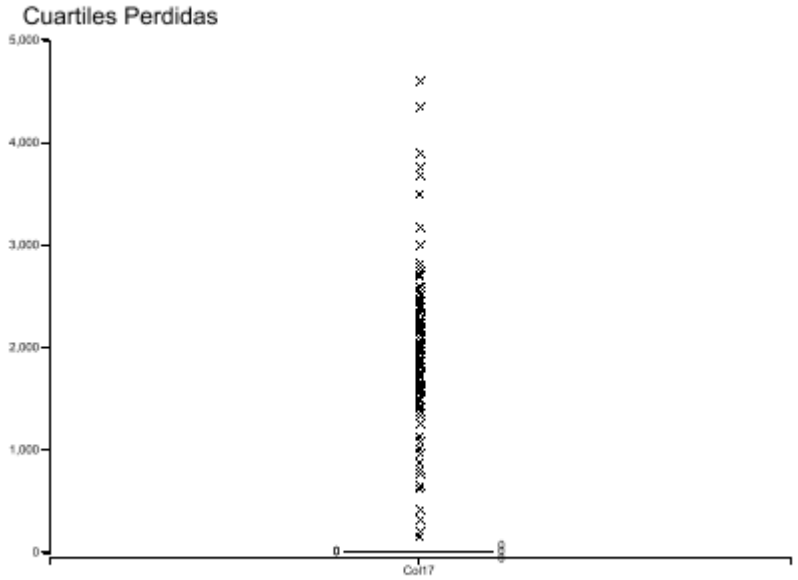
Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

CAPLOSS		
Dato	Valor	Interpretación
Mínimo	0	Minimo capital perdido.
Máximo	4608	Máximo capital perdido.
Media	37.314	Capital promedio perdido.
Desviación estándar	271.896	
Varianza	73,927.668	

Cuartiles



Descripción

Para el capital perdido tenemos que el minimo es 0 y el maximo es de 4,608 dolares.
Por otro lado, en la imagen, podemos observar que los datos se encuentran dispersos, no podemos notar un intervalo definido donde se concentren la mayoria de los datos. En la parte baja notamos un pequeño lugar donde en efecto se concentra la mayoria, confirmando que la media ronda los 37 dolares.

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

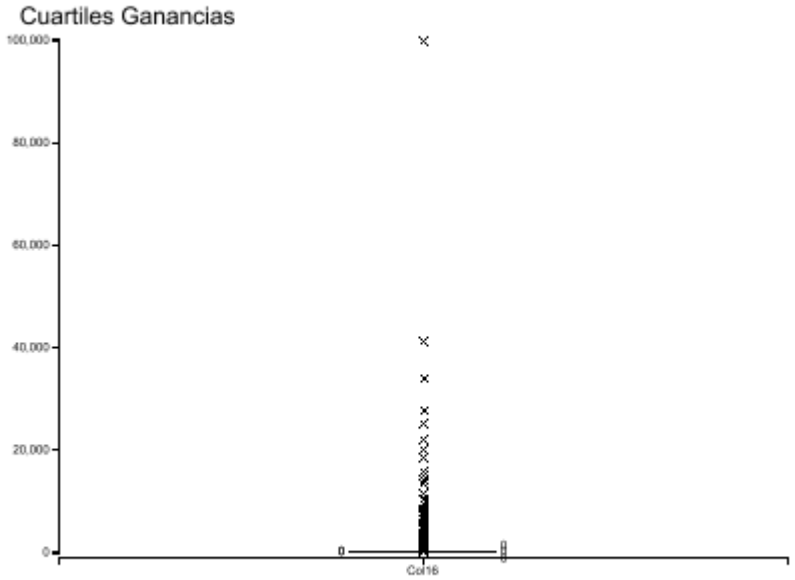
Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

CAPGAIN		
Dato	Valor	Interpretación
Mínimo	0	Minimo capital ganado.
Máximo	99,999	Máximo capital ganado.
Media	197.53	Capital promedio ganado.
Desviación estándar	1,984.164	
Varianza	3,936,905.423	

Cuartiles



Descripción

En el capital ganado tenemos que lo minimo es de 0 dolares y el maximo es de 99,999 dolares. Al ser un rango más amplio la desviacion es mayor, esto tambien lo podemos visualizar en la grafica puesto que tenemos muchos datos atipicos.

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

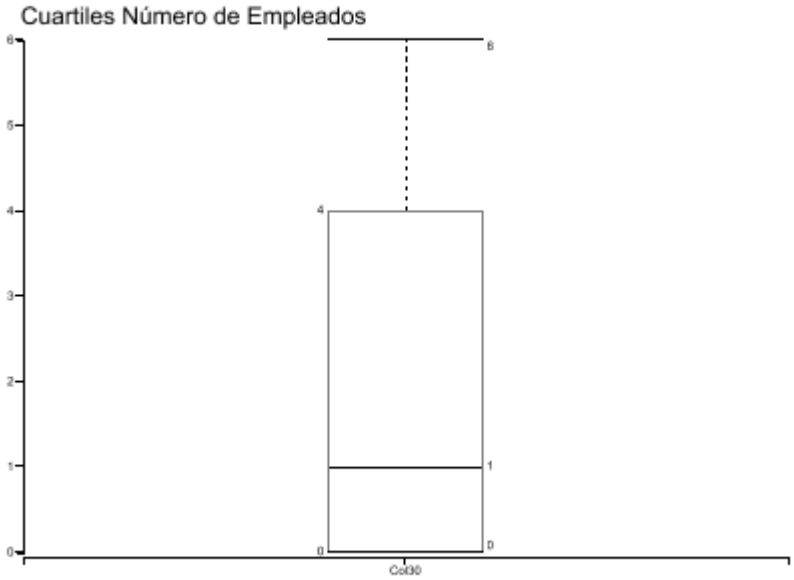
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

NOEMP		
Dato	Valor	Interpretación
Mínimo	0	Minimo de empleados.
Máximo	6	Máximo de empleados.
Media	1.952	Empleados promedio.
Desviación estándar	0.554	
Varianza	0.37	
Cuartiles		



Descripción	Este atributo hace referencia a la cantidad de empleados que se tienen, vemos que el minimo es de 0 empleados y el maximo de 6 empleados. Resultando entonces que la media es de dos empleados. En el grafico, podemos notar que el 75% tiene menos de cuatro empleados.

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

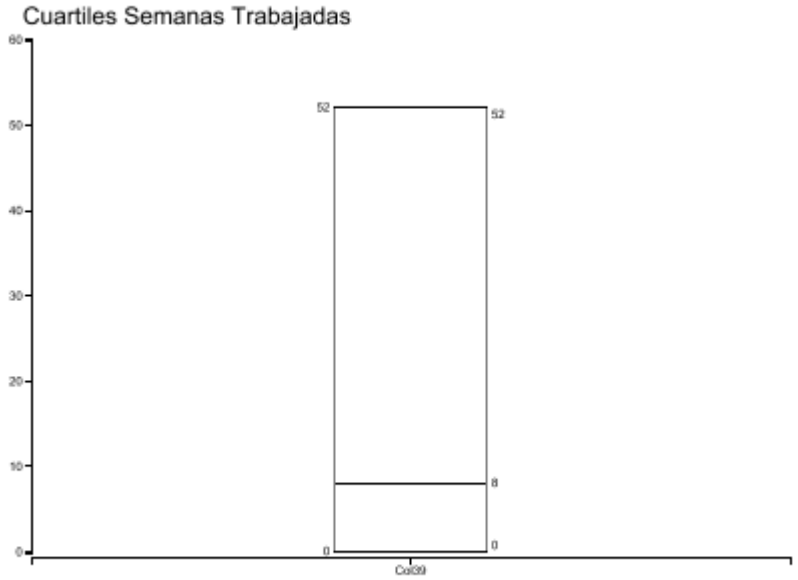
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

WKSWORK		
Dato	Valor	Interpretación
Mínimo	0	Mínimo número de semanas trabajadas
Máximo	52	Máximo número de semanas trabajadas
Media	23.17	Número promedio de semanas trabajadas
Desviación estándar	24.41	
Varianza	595.92	
Cuartiles		



Descripción	Las semanas trabajadas tienen un rango de 0 a 52 semanas, el maximo refiriendose entonces a que se trabajo un año entero. El grafico nos arroja que el 50% de las personas que formaron parte de esto reportaron haber trabajado menos de 8 semanas, el otro 50% trabajo entre 8 y 52 semanas.

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional
Escuela Superior de Computo
Trabajadores en Estados Unidos

Tablas de datos descriptivas

En esta seccion se presentaran las tablas de datos descriptivas, en ellas podremos ver la frecuencia acumulada, la segmentación de archivos (binner) y descripción de datos por segmentos.

Tabla de frecuencia de edades (AAGE)

Edades	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
0 a 4 años	15,810	15,810	7.92%	7.92%
5 a 19 años	45,085	60,895	22.59%	30.51%
20 a 29 años	27,239	82,909	11.03%	41.54%
30 a 44 años	47,991	120,721	18.95%	60.49%
45 a 64 años	49,109	168,356	23.87%	84.36%
65 a 89 años	23,584	198,798	15.25%	99.64%
90 o más años	725	199,523	0.36%	100%

Tabla de frecuencia de tipo de trabajador (ACLSWKR)

Tipo de trabajador	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Federal government	2,925	2,925	1.41%	1.41%
Local government	7,784	10,709	3.90%	5.31%
Never worked	439	11,148	0.22%	5.53%
Not in the universe	100,245	111,393	50.24%	55.77%
Private	72,028	183,421	36.10%	91.87%
Self-employed incorporated	3,265	186,686	1.63%	93.5%
Self imployed not incorporated	8,445	195,131	4.23%	97.73%
State government	4,227	199,358	2.11%	99.84%
Without pay	165	199,523	0.08%	99.97%

Tabla de frecuencia de nivel estudios (AHGA)

Tipo de estudios	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Children	47,422	47,422	23.76%	23.76%
7th and 8th grade	8,007	55,429	4.01%	27.77%
9th grade	6,230	61,659	3.12%	30.89%
10th grade	7,557	69,216	3.78%	34.67%
High school graduate	48,407	117,623	24.26%	58.93%
11th grade	6,876	124,499	3.44%	62.37%
12th grade no diploma	2,126	126,625	1.06%	63.43%
5th or 6th grade	3,277	129,902	1.64%	65.07%
Less than 1st grade	819	130,721	0.41%	65.48%

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena
Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Tipo de estudios	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Bachelors degree(BA AB BS)	19,865	150,586	9.95%	75.43%
1st 2nd 3rd or 4th grade	1,799	152,385	0.90%	76.33%
Some college but no degree	27,820	180,205	13.94%	90.27%
Masters degree(MA MS MEng MEd MSW MBA)	6,541	186,746	3.27%	93.54%
Associates degree-occup /vocational	5,358	192,104	2.68%	96.22%
Associates degree-academic program	4,363	196,467	2.18%	98.4%
Doctorate degree(PhD EdD)	1,263	197,730	0.63%	99.03%
Prof school degree (MD DDS DVM LLB JD)	1,793	199,523	0.89%	99.93%

Tabla de frecuencia de salario por hora (AHRSPAY)

Salario	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
0 a 1,666 dólares	198,308	198,308	99.39%	99.39%
1,666 a 3,333 dólares	1,132	199,440	0.56%	99.95%
3,333 a 4,999 dólares	38	199,478	0.019%	99.96%
4,999 a 6,666 dólares	24	199,502	0.012%	99.98%
6,666 a 8,832 dólares	11	199,513	0.005%	99.98%
8,832 a 9,999 dólares	10	199,523	0.005%	99.99%

Tabla de frecuencia de educacion reciente (AHSCOL)

Tipo de educacion	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
College or University	5,688	5,688	2.85%	2.85%
High school	6,892	12,580	3.45%	6.30%
Not in the universe	186,943	199,523	93.69%	99.99%

Tabla de frecuencia de estado marital (AMARITL)

Estado	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Never married	86,485	86,485	43.34%	43.34%
Married-civilian spouse present	84,222	170,707	42.21%	85.55%
Divorced	12,710	183,417	6.37%	91.92%
Widowed	10,463	193,880	5.24%	97.16%
Separated	3,460	197,340	1.73%	98.89%
Married-spouse absent	1,518	198,858	0.76%	99.65%

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Estado	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Married-A F spouse present	665	199,523	0.33%	99.98%

Tabla de frecuencia de raza (ARACE)

Raza	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
White	167,365	167,365	83.88%	83.88%
Black	20,415	187,780	10.23%	94.11%
Asian or Pacific Islander	5,835	193,615	2.92%	97.03%
Amer Indian Aleut or Eskimo	2,251	195,866	1.12%	98.15%
Other	3,657	199,523	1.83%	99.98%

Tabla de frecuencia de capital ganado (CAPGAIN)

Capital	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
0 a 24,999 dólares	198,985	198,985	99.73%	99.73%
24,999 a 49,999 dólares	148	198,647	0.07%	99.80%
49,999 a 74,999 dólares	0	198,647	0.0%	99.80%
74,999 a 99,999 dólares	390	199,523	0.19%	99.99%

Tabla de frecuencia de capital perdido (CAPLOSS)

Capital	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
0 a 768 dólares	195,699	195,699	98.08%	98.08%
768 a 1,536 dólares	332	196,031	0.16%	98.24%
1,536 a 2,304 dólares	2,946	198,977	1.47%	99.71%
2,304 a 3,072 dólares	483	199,460	0.24%	99.95%
3,072 a 3,840 dólares	27	199,487	0.017%	99.96%
3,840 a 4,608 dólares	36	199,523	0.018%	99.99%

Tabla de frecuencia de numero de empleados(NOEMP)

# empleados	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Menos de 2 empleados	119,092	119,092	59.68%	59.68%
De 2 a 4 empleados	23,506	142,598	11.78%	71.46%
De 4 a 6 empleados	56,925	199,523	28.53%	99.99%

Tabla de frecuencia de semanas trabajadas(WKSWORK)

Semanas	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Menos de 13 semanas	103,103	103,103	51.67%	51.67%
Entre 13 y 26 semanas	5,828	108,931	2.92%	54.59%
Entre 26 y 39 semanas	7,913	116,844	3.96%	58.55%

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

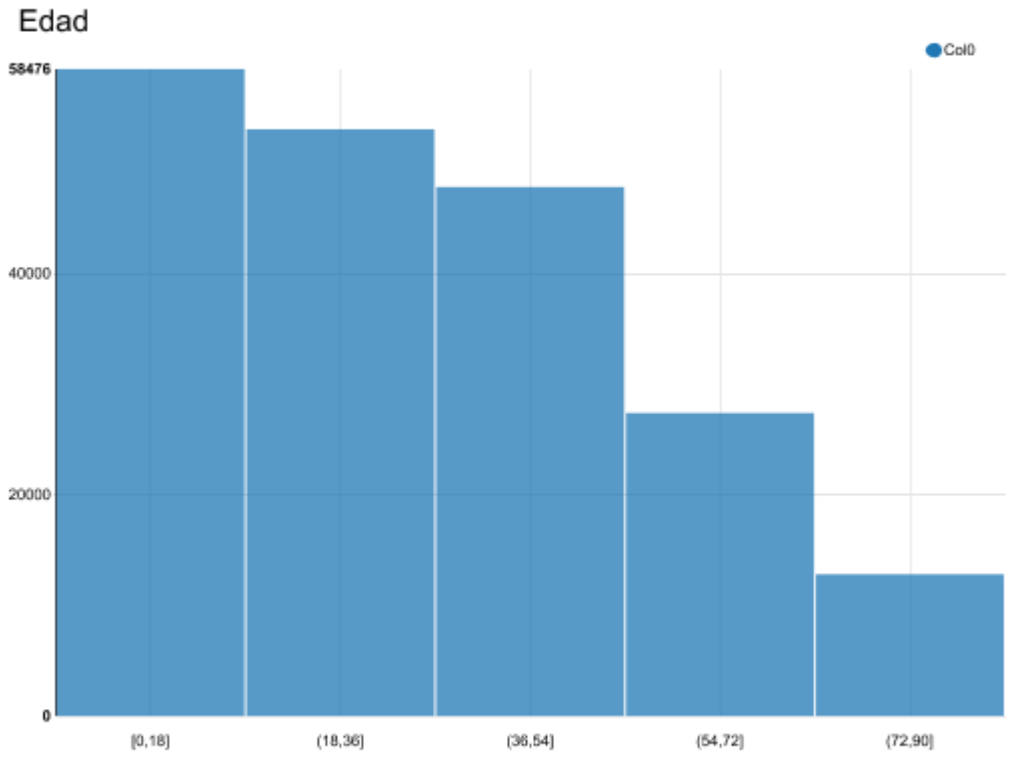
Semanas	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Entre 39 y 52 semanas	82,679	199,523	41.43%	99.98%

Tabla de frecuencia de año de recolección (YEAR)

Año	Frecuencia	Frecuencia acumulada	% de frecuencia	% f. acumulada
Año 94	99,827	99,827	50.03%	50.03%
Año 95	99,696	199,523	49.96%	100%

Representaciones Graficas

Dato	Grafico
AAGE	



Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

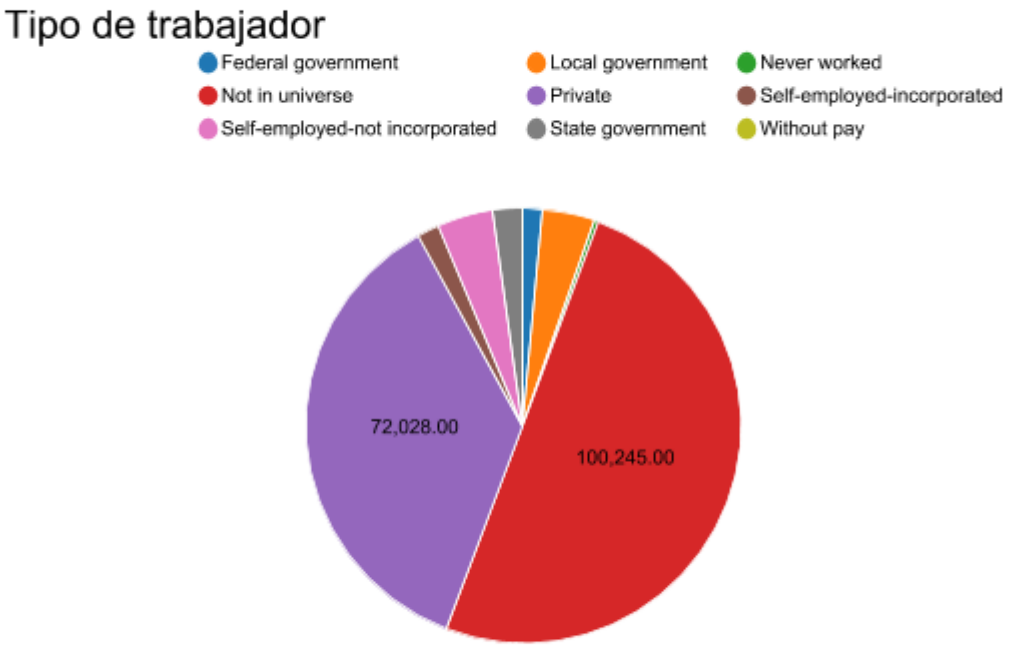
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Dato	Grafico
ACLSWKR	



Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

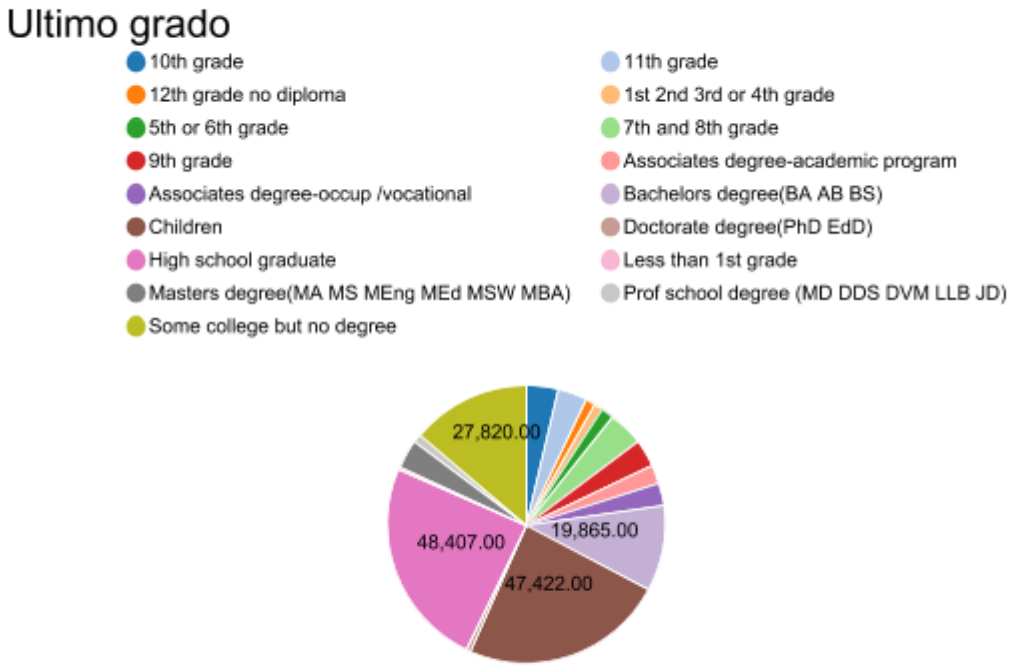
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Dato	Grafico
AHGA	

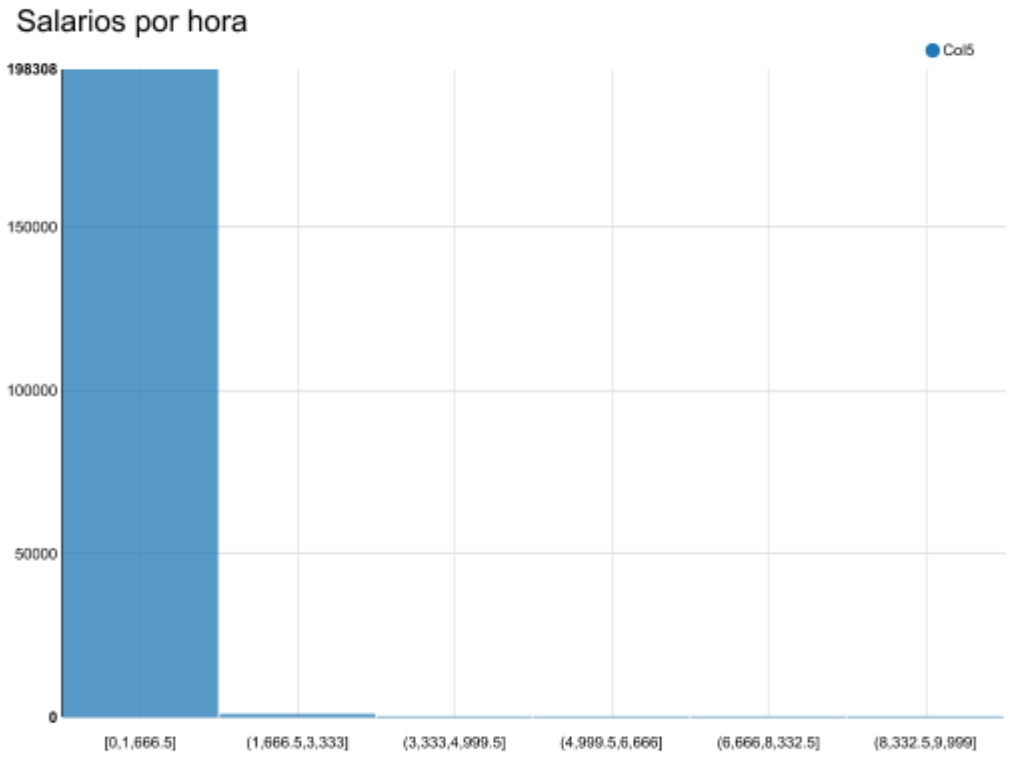


Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena
Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Dato	Grafico
AHRSPAY	

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

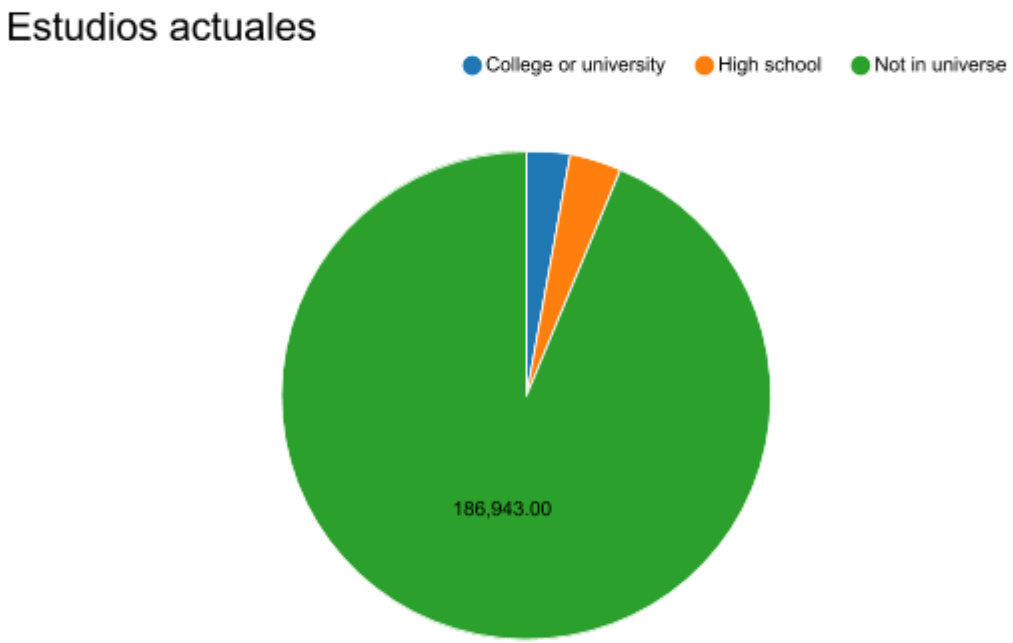
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Dato	Grafico
AHSCOL	



Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

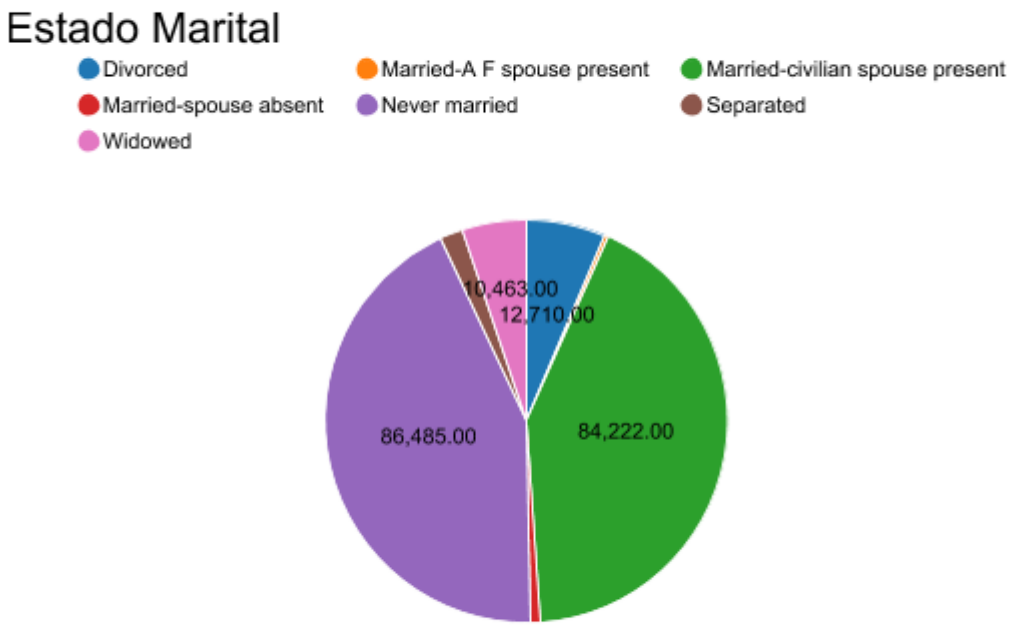
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Dato	Grafico
AMARITL	



Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

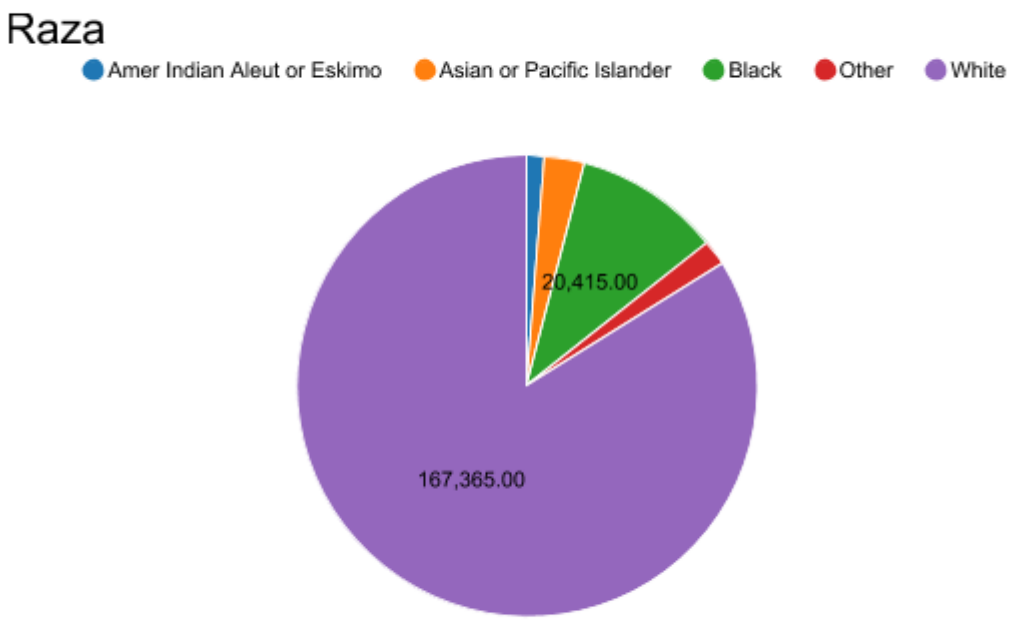
Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Dato	Grafico
ARACE	

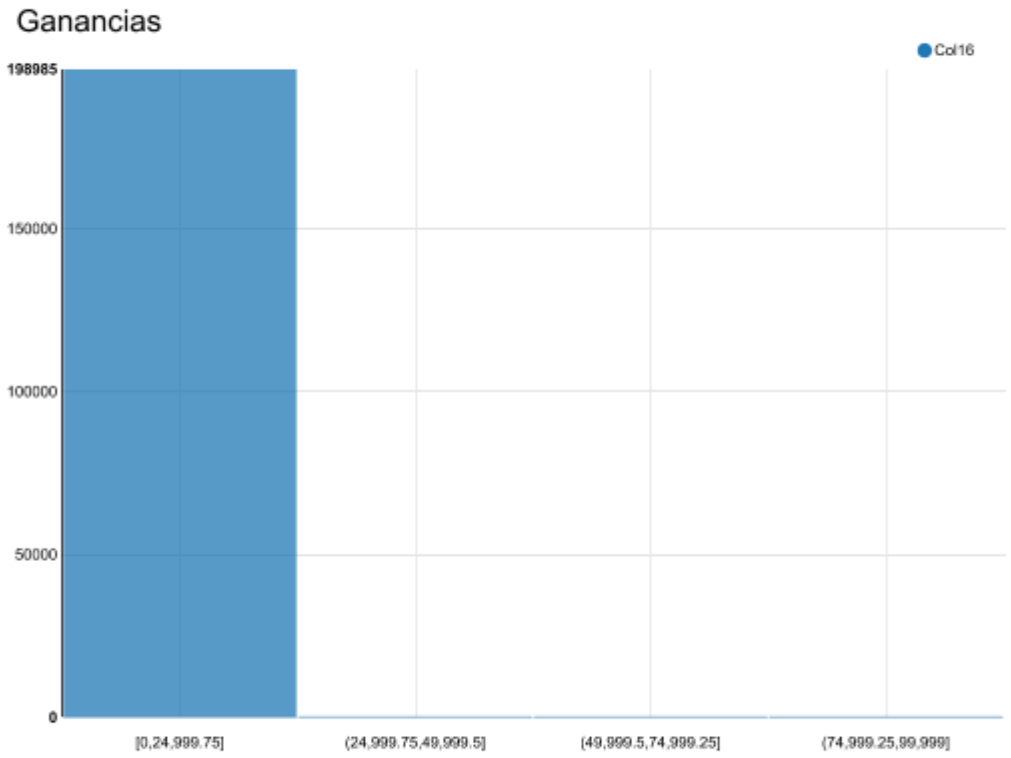
Data Mining
3CV11

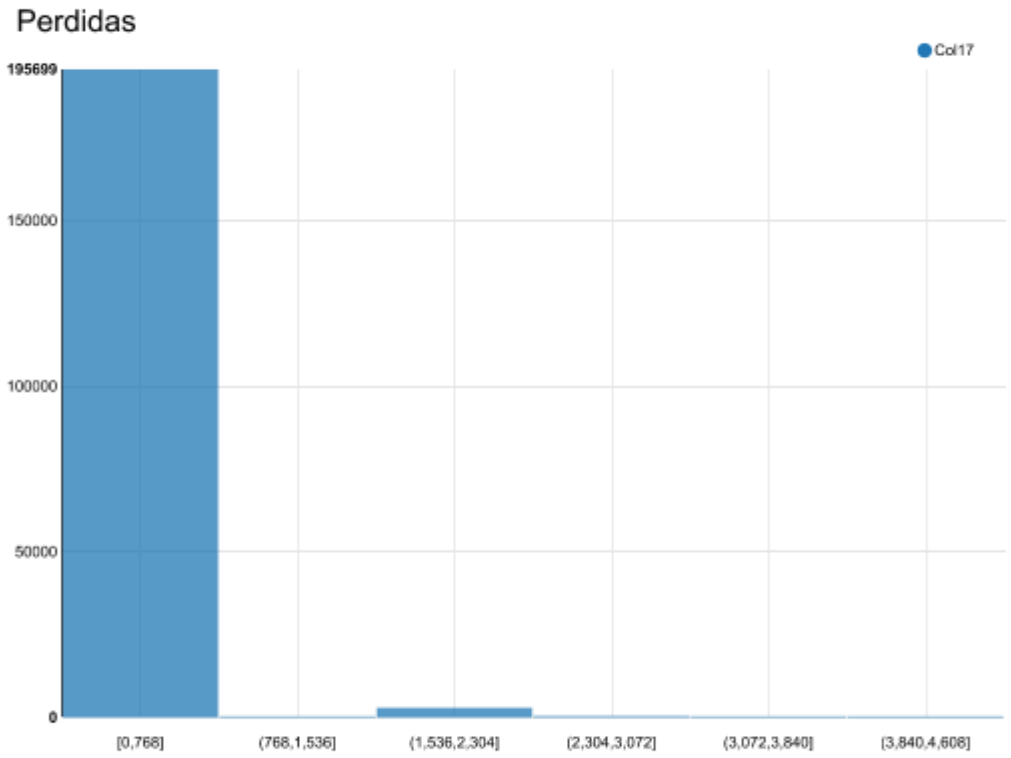
Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Dato	Grafico
CAPGAIN	



Dato	Grafico
CAPLOSS	

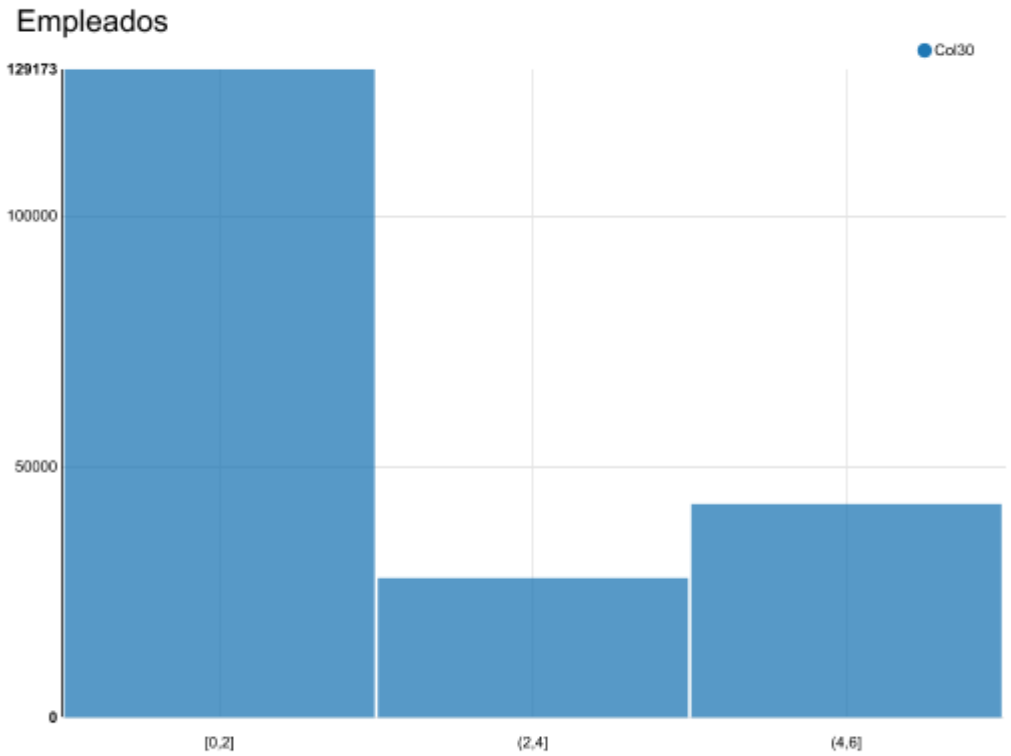
Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Dato	Grafico
NOEMP	



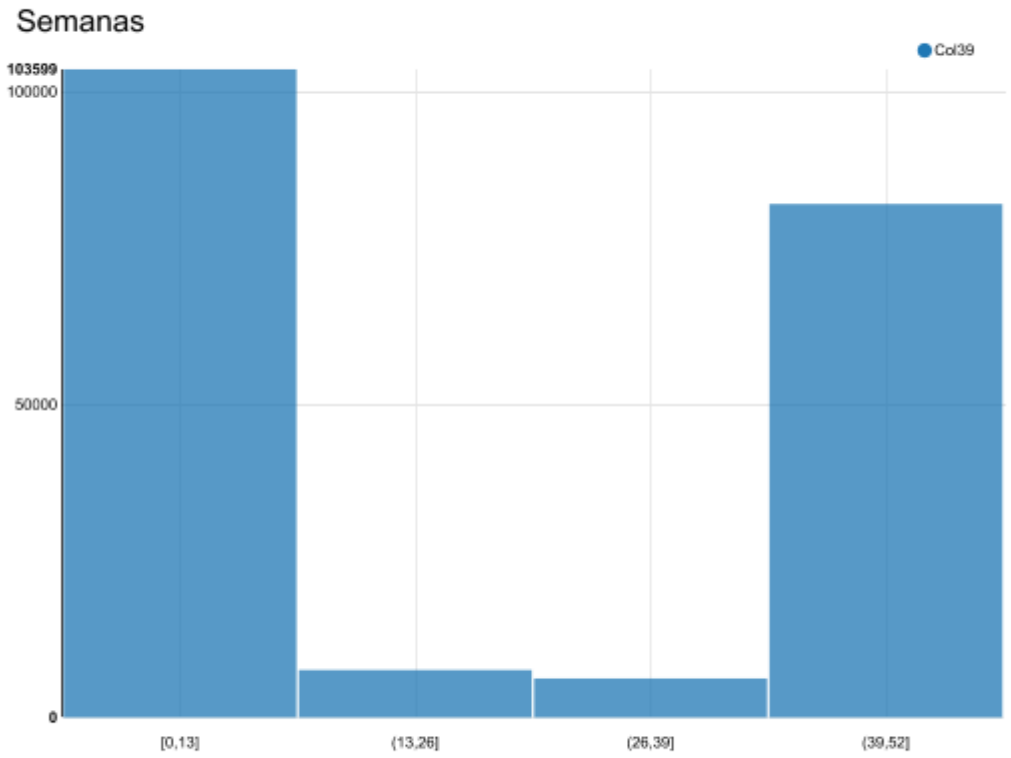
Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Dato	Grafico
WKSWORK	



Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

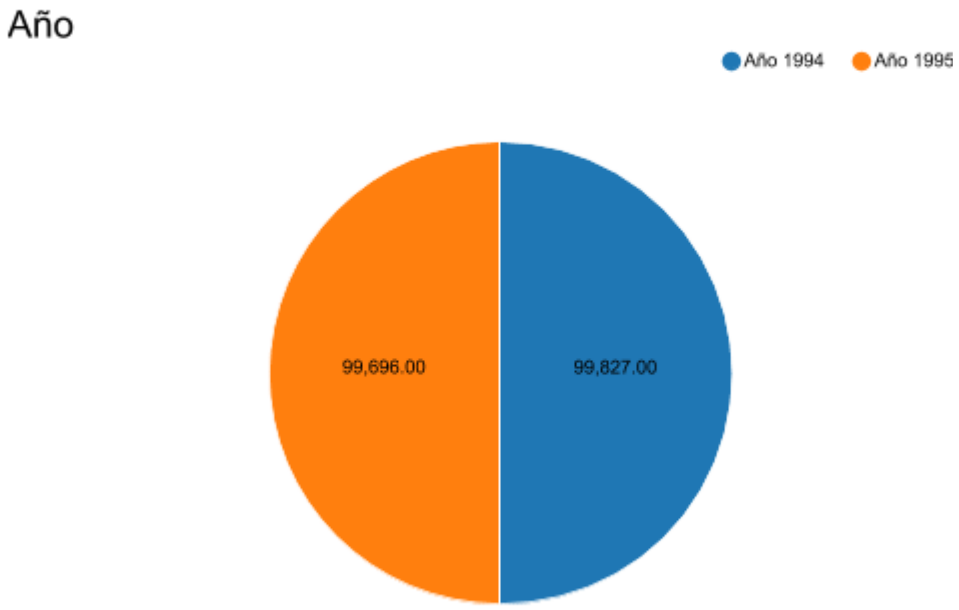
15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Dato	Grafico
YEAR	



Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional
Escuela Superior de Computo
Trabajadores en Estados Unidos

Tratamiento de datos

Imputación: Los métodos de imputación consisten en estimar los valores ausentes en base a los valores válidos de otras variables y/o casos de la muestra. En este caso como estos atributos si tiene datos faltantes, optamos por utilizar este método. Columnas 25 a 27, dado que estamos trabajando con datos de tipo nominal, se rellenarán los datos faltantes con el dato que más se repita, este tratamiento se nos ayuda no solo a dejar de tener inconsistencias en los datos que nos impidan llegar a conclusiones o generar conclusiones precisas, ademas de que el rellenar los datos faltantes con el que mas se repite nos reduce la posibilidad de equivocarnos y aproximarnos a un resultado mas certero y cernano a la realidad, ya que no podemos evitar los errores humanos, pero podemos minimizar la perdida de datos y la restablecer la consistencia de los datos.

File Table - 0:1 - File Reader

File Edit Hilite Navigation View

Table "census-income.data" - Rows: 199523 Spec - Columns: 42 Properties Flow Variables

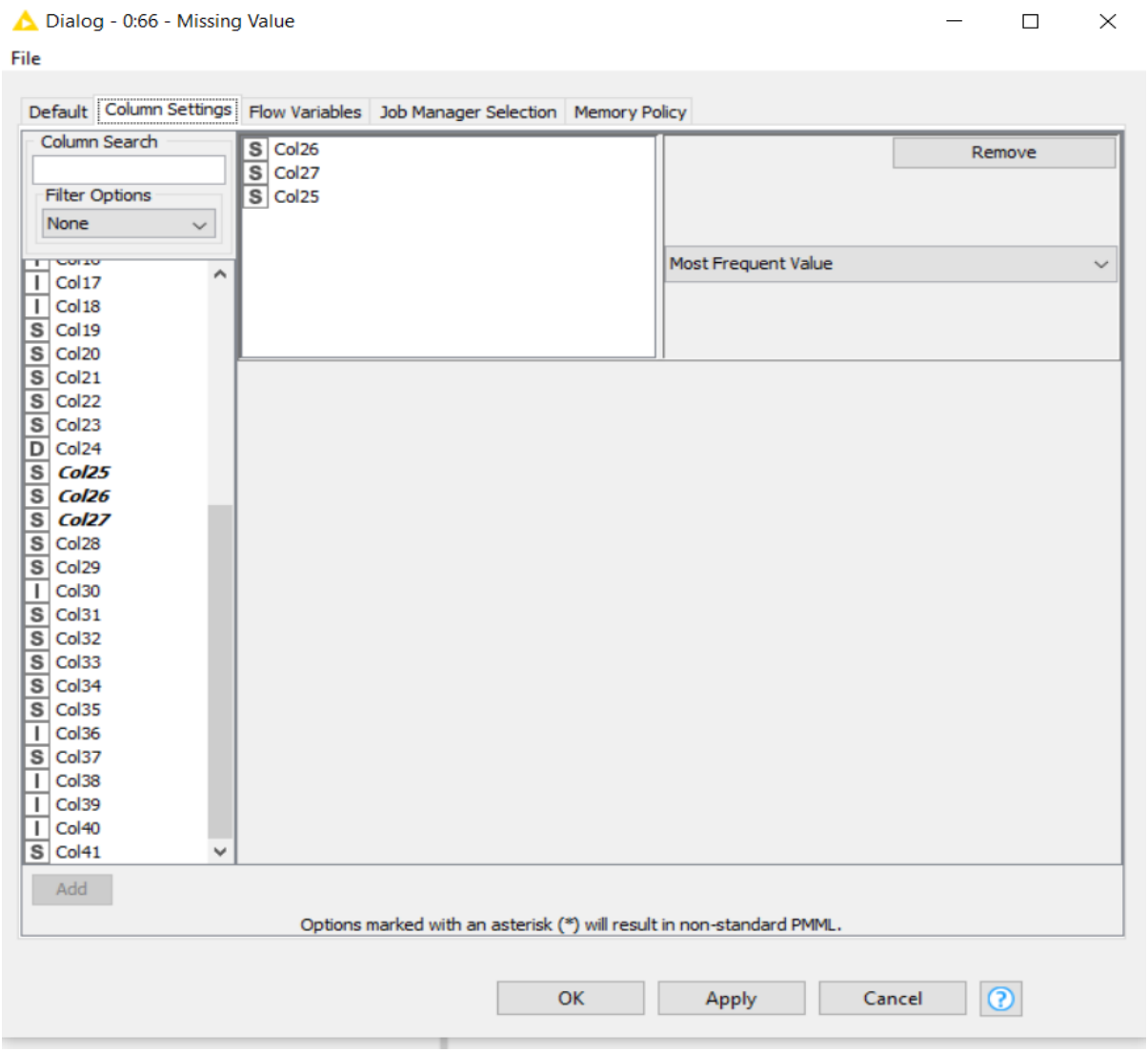
Row ID	Col21	Col22	Col23	Col24	Col25	Col26	Col27	Col28	Col29	Col30	Col31	S
Row96	Not in universe	Householder	Householder	958.8	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	Unit
Row97	Not in universe	Spouse of householder	Spouse of householder	1,897.64	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	?
Row98	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	1,353.41	?	?	?	Not in universe under 1 year old	?	0	Both parents...	Unit
Row99	Not in universe	Child 18+ never marr Not in a subfamily	Child 18 or older	996.16	?	?	?	Not in universe under 1 year old	?	6	Not in universe	Unit
Row100	Not in universe	Householder	Householder	2,634.31	?	?	?	Not in universe under 1 year old	?	0	Not in universe	Unit
Row101	Not in universe	Spouse of householder	Spouse of householder	1,152.47	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	Unit
Row102	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	1,117.33	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Both parents...	Unit
Row103	Not in universe	Spouse of householder	Spouse of householder	636.85	Nonmover	Nonmover	Nonmover	Yes	Not in universe	2	Not in universe	Unit
Row104	Not in universe	Householder	Householder	2,730.77	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Not in universe	Unit
Row105	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	1,428.97	Not in universe	Not in universe	Not in universe	Not in universe under 1 year old	Not in universe	0	Both parents...	Mex
Row106	Not in universe	Child 18+ never marr Not in a subfamily	Child 18 or older	2,147.49	?	?	?	Not in universe under 1 year old	?	3	Not in universe	Unit
Row107	Alaska	Child <18 never marr not in subfamily	Child under 18 never married	215.84	MSA to MSA	Same county	Same county	No	No	0	Both parents...	Unit
Row108	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	1,289.56	?	?	?	Not in universe under 1 year old	?	0	Both parents...	Mex
Row109	New Mexico	Secondary individual	Nonrelative of householder	1,208.48	MSA to MSA	Different div...	Different st...	No	No	0	Not in universe	Unit
Row110	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	454.07	?	?	?	Not in universe under 1 year old	?	0	Mother only ...	Unit
Row111	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	2,887.5	?	?	?	Not in universe under 1 year old	?	0	Both parents...	Mex
Row112	Not in universe	Nonfamily householder	Householder	1,103.79	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	Unit
Row113	Not in universe	Householder	Householder	573.79	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Not in universe	Unit
Row114	Not in universe	Householder	Householder	1,516.05	?	?	?	Not in universe under 1 year old	?	0	Not in universe	Unit
Row115	Not in universe	Householder	Householder	1,282.49	Nonmover	Nonmover	Nonmover	Yes	Not in universe	1	Not in universe	Dom
Row116	Not in universe	Spouse of householder	Spouse of householder	2,700.08	Nonmover	Nonmover	Nonmover	Yes	Not in universe	4	Not in universe	Unit
Row117	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	1,570.8	?	?	?	Not in universe under 1 year old	?	0	Both parents...	Mex
Row118	Not in universe	Householder	Householder	272.14	Nonmover	Nonmover	Nonmover	Yes	Not in universe	3	Not in universe	El-S
Row119	Not in universe	Householder	Householder	2,172.64	?	?	?	Not in universe under 1 year old	?	0	Not in universe	Unit
Row120	Nevada	Nonfamily householder	Householder	803.77	MSA to MSA	Same county	Same county	No	No	3	Not in universe	Unit
Row121	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	709.9	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Both parents...	Unit
Row122	Not in universe	Spouse of householder	Spouse of householder	3,497.86	?	?	?	Not in universe under 1 year old	?	3	Not in universe	Can
Row123	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	1,905.6	?	?	?	Not in universe under 1 year old	?	0	Both parents...	Mex
Row124	Not in universe	Nonfamily householder	Householder	2,117.38	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	Sco
Row125	Not in universe	Child <18 never marr not in subfamily	Child under 18 never married	2,288.5	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Both parents...	Mex
Row126	Not in universe	Nonfamily householder	Householder	359.51	?	?	?	Not in universe under 1 year old	?	1	Not in universe	Unit
Row127	Not in universe	Child 18+ never marr Not in a subfamily	Child 18 or older	1,680.95	Nonmover	Nonmover	Nonmover	Yes	Not in universe	1	Not in universe	Unit
Row128	Not in universe	Child 18+ never marr Not in a subfamily	Child 18 or older	1,782.93	?	?	?	Not in universe under 1 year old	?	4	Not in universe	Unit
Row129	Not in universe	Nonfamily householder	Householder	1,989.51	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Not in universe	?

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena
Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021



Así como resultado del método de imputación se obtiene:

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Instituto Politecnico Nacional

Escuela Superior de Computo

Trabajadores en Estados Unidos

Output table - 0:66 - Missing Value

File Edit Hilite Navigation View

Table "default" - Rows: 199523 | Spec - Columns: 42 | Properties | Flow Variables

Row ID		S Col23	D Col24	S Col25	S Col26	S Col27	S Col28	S Col29	D Col30	S Col31	S Col32	S Col33	S
Row0	ever marr not in subfamily	Other relative of householder	1,700.09	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	0	Not in universe	United-States	United-States	Un
Row1		Householder	1,053.55	MSA to MSA	Same county	Same county	No	Yes	1	Not in universe	United-States	United-States	Un
Row2	er marr Not in a subfamily	Child 18 or older	991.95	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	0	Not in universe	Vietnam	Vietnam	Vie
Row3	er marr not in subfamily	Child under 18 never married	1,758.14	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Both parents...	United-States	United-States	Un
Row4	er marr not in subfamily	Child under 18 never married	1,069.16	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Both parents...	United-States	United-States	Un
Row5	seholder	Spouse of householder	162.61	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	1	Not in universe	Philippines	United-States	Un
Row6		Householder	1,535.86	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	United-States	United-States	Un
Row7	vidual	Nonrelative of householder	898.83	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	4	Not in universe	United-States	United-States	Un
Row8	seholder	Spouse of householder	1,661.53	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	5	Not in universe	United-States	United-States	Un
Row9		Householder	1,146.79	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	United-States	United-States	Un
Row10	er marr not in subfamily	Child under 18 never married	2,466.24	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Both parents...	United-States	United-States	Un
Row11	never marr not in subfa...	Other relative of householder	2,021.27	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	0	Not in universe	United-States	United-States	Un
Row12		Householder	2,441.22	Nonmover	Nonmover	Nonmover	Yes	Not in universe	3	Not in universe	United-States	United-States	Un
Row13		Householder	978.16	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	Columbia	Columbia	Co
Row14	seholder	Householder	2,604.91	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	6	Not in universe	United-States	United-States	Un
Row15	er marr not in subfamily	Child under 18 never married	1,520.08	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Both parents...	United-States	United-States	Un
Row16	seholder	Householder	404.9	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	6	Not in universe	Germany	United-States	Un
Row17	seholder	Spouse of householder	1,274.04	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Not in universe	Mexico	Mexico	Me
Row18	er marr not in subfamily	Child under 18 never married	1,555.29	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	0	Both parents...	United-States	El-Salvador	Un
Row19		Householder	1,790.75	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	4	Not in universe	United-States	United-States	Un
Row20	er marr not in subfamily	Child under 18 never married	455.02	MSA to MSA	Different re...	Different st...	No	Yes	0	Both parents...	United-States	United-States	Un
Row21		Householder	1,004.69	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	United-States	United-States	Un
Row22	seholder	Spouse of householder	1,500.08	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	2	Not in universe	United-States	United-States	Un
Row23		Householder	999.46	Nonmover	Nonmover	Nonmover	Yes	Not in universe	1	Not in universe	United-States	United-States	Un
Row24		Householder	1,483.69	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Not in universe	Japan	United-States	Un
Row25	er marr not in subfamily	Child under 18 never married	1,660.53	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Both parents...	United-States	United-States	Un
Row26	er marr not in subfamily	Child under 18 never married	848.25	MSA to MSA	Different co...	Different co...	No	No	0	Mother only ...	United-States	United-States	Un
Row27	seholder	Spouse of householder	2,671.99	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	3	Not in universe	United-States	United-States	Un
Row28	er marr not in subfamily	Child under 18 never married	1,188.42	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	0	Both parents...	United-States	United-States	Un
Row29	vidual	Nonrelative of householder	1,331.35	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	United-States	United-States	Un
Row30		Householder	711.15	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	6	Not in universe	Mexico	Mexico	Me
Row31	seholder	Spouse of householder	1,578.65	Nonmover	Nonmover	Nonmover	Not in universe under 1 year old	Not in universe	6	Not in universe	United-States	United-States	Un
Row32		Householder	1,629.02	Nonmover	Nonmover	Nonmover	Yes	Not in universe	1	Not in universe	United-States	United-States	Un
Row33	er marr Not in a subfamily	Child 18 or older	1,998.03	Nonmover	Nonmover	Nonmover	Yes	Not in universe	1	Not in universe	United-States	United-States	Un
Row34	seholder	Householder	463.55	Nonmover	Nonmover	Nonmover	Yes	Not in universe	0	Not in universe	United-States	United-States	Un
Row35		Householder	2,492.74	Nonmover	Nonmover	Nonmover	Yes	Not in universe	3	Not in universe	Peru	Peru	Pe
Row36		Householder	980.1	Nonmover	Nonmover	Nonmover	Yes	Not in universe	6	Not in universe	United-States	United-States	Un

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021

Conclusiones

Conclusión Arévalo Andrade Miguel Ángel: Con este proyecto se puso de manifiesto la importancia y las grandes cosas que se pueden hacer con la información obtenida en el mundo, en las personas, en este caso en la gente trabajadora de Estados Unidos y los métodos de tratamiento de datos para eliminar inconsistencias que pueda haber en los data set. La limpieza de datos también es importante porque mejora la calidad de los datos y, al hacerlo, aumenta la productividad general. Cuando limpia los datos, toda la información desactualizada o incorrecta desaparece, dejándolo con información de la más alta calidad.

Conclusión Aguilar Martínez Oswaldo: En muchas ocasiones tenemos conjuntos de datos enormes, que pensamos que no sirven o no contienen absolutamente nada de relevancia o valor, pero si los conjuntos que nosotros mismos generamos pueden llegar a tener informacion que con los tratamientos adecuados nos brindan mucha mas informacion y no solo con respecto al objetivo por el que se generaron, en el desarrollo de esta practica encontramos un banco de datos muy grande que brindaba informacion acerca de todo, tenia un poco de todo, desde la situacion de ciertas personas, la edad, los ingresos incluso de sus origenes, cuando vemos el cumulto de datos que es muy grande, llegamos a pensar que no tienen mucha relevancia, pero una vez tratada, obtenemos datos y estadisticas, de los aspectos ya mencionados, sin duda el tratar los datos nos ayuda a darle un nuevo uso a los datos y no solo generarlos sin ningun objetivo.

Conclusión Guerrero Espinosa Ximena: A mi parecer, en general muchas personas(incluyéndome), tenemos una idea errónea de lo que es la minería de datos, así como cuál es su propósito. Tenemos una vaga noción del tema, pero no comprendemos por completo el concepto mismo por lo que creo firmemente que este proyecto me ayudo a aterrizar los conceptos básicos relacionados a la minería de datos. De la misma manera fui capaz de, por mis propios medios, imaginar cuales serían las posibles aplicaciones para esto y quitarme de la cabeza la idea de que es un proceso complicado.
Enfocándolo al conjunto de datos que seleccionamos me parece que, tras la limpieza, y el tratamiento que le dimos nos ayudo a entender de mejor manera lo que representaban todo este conjunto, que, a simple vista, podrían solo ser palabras al azar. En nuestro conjunto de datos, por ejemplo, las columnas no tenían nombre por lo que el proceso de limpieza resulto un poco más complicado, pero como mencione me ayudo a quitarme el miedo a la herramienta de Knime.

Data Mining
3CV11

Integrantes
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel
Guerrero Espinosa Ximena
Mariana

Profesora Ocampo Botello Fabiola

15 de Abril, 2021