ITESO

DEPARTAMENTO DE MATEMÁTICAS Y FÍSICA

Asignatura: Ciencia de Datos e Inteligencia de Negocios

EXAMEN (Medición Estadística de Datos y Medidas de Similitud)

Nombre: Oscar Eduardo Flores Hernández

Lea detenidamente los reactivos y responda con claridad. Si se requiere hacer uso de hojas impresas para la realización de cálculos, es necesario que se adjunten en forma digital cuando se haga entrega del examen.

1. (2 puntos) Explique cuáles son y cuáles son sus características de los diferentes tipos de datos que podemos encontrar.

   Existen 5 tipos de datos:

   - Binarios; Están compuestos por datos con únicamente dos opciones (verdadero/falso), (si/no), (1/0), ... Representan una dualidad y se utilizan métricas como Jaccard y emparejamiento simple para encontrar similaridades entre ellos.

   - Nominales; Los nominales son aquellos que pueden tener varios estados y no necesariamente afecta el orden de los factores. El caso más representativo sería decir que los colores pueden ser representados como (azul = 1, rojo = 2, negro = 3, ...), son muy similares a los binarios en cuanto a los calculos a ejecutar, sin embargo, para llegar a ser calculados se les tiene que convertir en variables ‘dummy’, que prácticamente cambian sus clasificaciones a binarias.

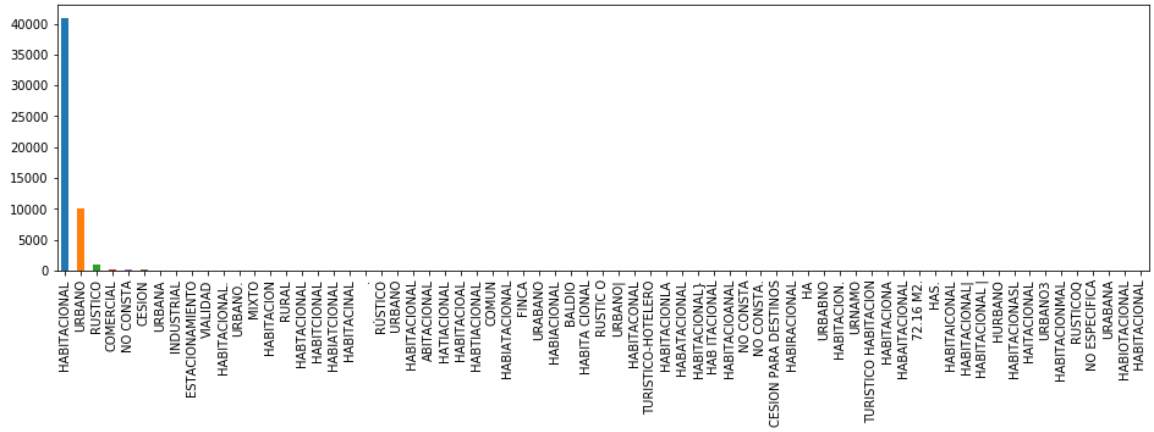
   - Ordinales; Muy similares a las nominales, pueden tener muchos valores distintos (multiestado), sin embargo, en estos sí importa el orden en el cuál son presentados. Como ejemplo se podría suponer la cantidad de cuartos en una casa (1,2,3, ... 7?) en el que claramente 3 cuartos son más que 2 cuartos. Su principal característica es que no se puede tener 1/2 cuarto.

   - Cuantitativos; Son Todos los números Reales. aplican decimales, implica un orden entre los números. Su cálculo se hace fundamentalmente con distancia euclideana y Cos().

   - Genéticos; Sabemos que existen los datos genéticos, no estando seguro de la aplicación de los mismos, ni del surgimiento de ellos.

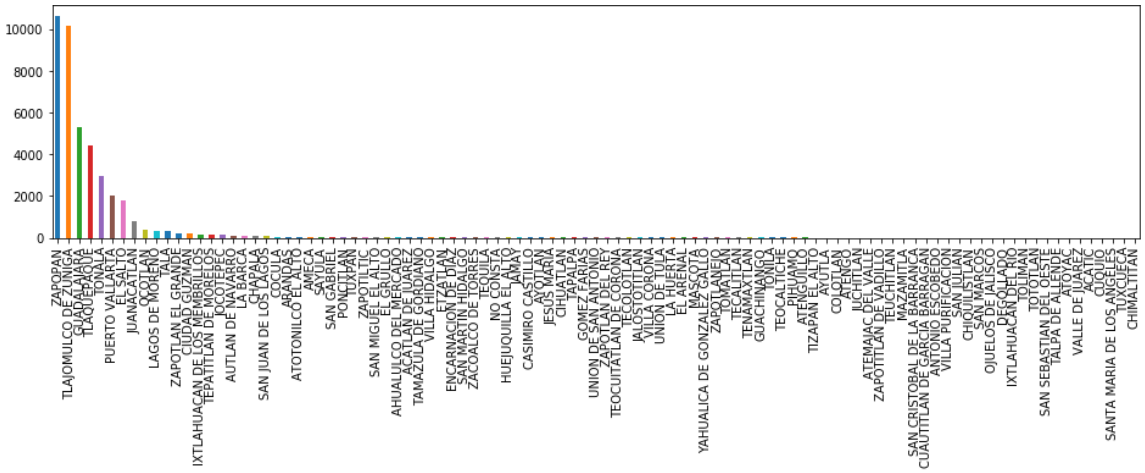
2. En la base de datos “cancelacion_2017.csv” contiene la lista de Inmuebles Registrados por Acto Jurídico en el Estado de Jalisco del 2017. Con los datos presentados en este archivo, responda las siguientes preguntas:

a. (1 punto) Determine qué tipos de uso de suelo son considerados en esta base de datos y genere un gráfico donde se muestre cuantos inmuebles se reportaron por cada tipo de uso de suelo.



Se puede notar que el principal uso de suelo es habitacional, así como también se puede notar la alta incidencia de errores en la escritura de la palabra misma. Se puede notar un caso partícular en el cual se escribe ‘72.16m2', no sabemos el uso de suelo, sin embargo parece que es la superficie construida en ese terreno partícularmente (dato que no es requerido).

b. (1 punto) Si consideramos solamente el uso de suelo “HABITACIONAL” determine cual es el municipio con menos inmuebles con este tipo de uso de suelo y que municipio tiene el mayor número de estos inmuebles.



Los tres municipios con más inmuebles son: Zapopan, Tlajomulco de Zuñiga, Guadalajara. Mientras que los que menos inmuebles tuvieron son Tuxcueca y Chimaltitlan. Parece muy lógico, a excepción del orden en los primeros tres; personalmente esperaba un orden Zapopan-Guadalajara-Tlajomulco, sin embargo, parece ser que Tlajomulco en esta tabla tiene registrados cerca de 3/2 la cantidad de inmuebles que Guadalajara tiene registrados.

3. Si se tuvieran dos bases de datos pequeñas como las siguientes:

Num	Trabajador	Estado Civil	Num Hijos	Ingreso
C1	Eduardo R.	Casado	1	50000.00
C2	Carlos Rojas	Divorciado	1	5000.50
C3	Aurora Flores	Casado	2	5000.00
C4	Saúl Carmona	Soltero	3	1000.00
C5	Lucía Morales	Divorciado	3	10000.00

Num	Municipio	Sector productivo	Población Actual (t 0 Las cantidades no corresponden a los valores reales.
C1	Zapopan	6	100,000.00
C2	Guadalajara	2	300,000.00
C3	Tlaquepaque	1	259,236.05
C4	Tequila	4	540,689.00
C5	Zapotlanejo	3	200,000.00

c. (1 punto) Indique cuantos tipos de variables hay en estas dos bases de datos. Haga una lista del nombre de variable y el tipo al que pertenece.

Se tienen 3 tipos de variables:

- Nominales; (Num, Trabajador y Estado Civil) y (Num, Municipio, Sector productivo)
- Ordinales; (Num Hijos) y (Nan)
- Cuantitativas; (Ingreso) y (Población Actual)

Nota: ¿Cómo se puede tener una población actual fraccionada? (Supongo que se debe a que las cuantitativas pueden tener decimales y se quería hacer esta distinción.)

d. (1 punto) Si se quisiera hacer un solo DataFrame de datos cuantitativos incluyendo todas variables de las dos tablas ¿Cuántas variables dummy habría que crear para el último DataFrame? Explique o justifique su respuesta.

se tendrían 28 variables Dummy: [5-Num,5-Trabajador,3-Estado_Civil]tabla1+[5-Num,5-Municipio,5-Sector_Productivo]tabla2. La variable Num se repite en ambos casos, sin embargo, se tiene que tomar como una Dummy diferente debido a que la clasificación del primero es distinta a la clasificación del segundo.

4. (2 puntos) Utiliza las mismas funciones para limpiar la base de datos de texto vista en clase para limpiar la base de datos llamada “dirty_Info_Alumnos_v2.csv”. El ejercicio consiste en normalizar toda la tabla teniendo encuentra las siguientes consideraciones:

✓ Dejar los nombres correctos en mayúsculas.

✓ Eliminar todos los caracteres especiales de la columna teléfono.

✓ Cuando no se tenga la longitud de 10 dígitos en el teléfono, se considerará el dato como “missing”.

✓ Dejar solo números en la columna semestre.

✓ Si el número de expediente no tiene 6 dígitos considerarlo como “missing”.

Agregue el código usado, la tabla generada y comente sus resultados.

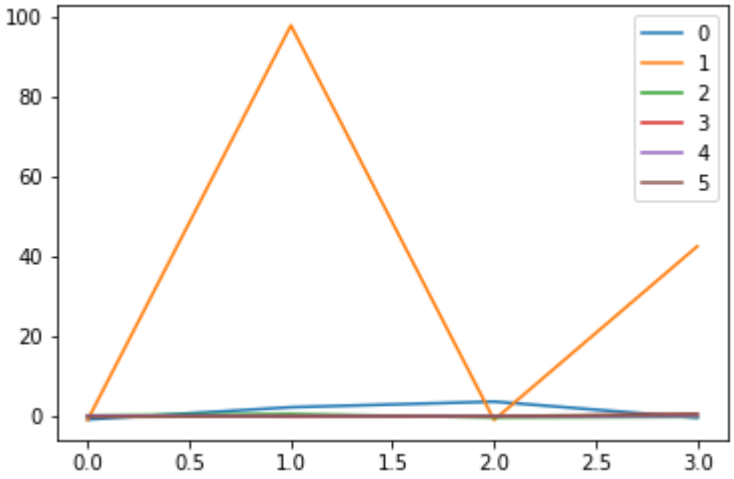
0	ALVAREZ DEL CASTILLO CASTAÑEDA JUAN MANUEL	20	13	639408	1062562979
1	ANAYA AVALOS VICTOR MAURICIO	20	7	930023	7354926370
2	ANSOLEAGA ALVAREZ MARIA FERNANDA	19	12	871375	7479367588
3	BARBA GALVAN SANTIAGO	19	1	439629	6589950295
4	BARBOZA ESPINOZA CARLOS ALFONSO	23	9	272305	2237277358
5	BUENO ARAGON PEDRO LUIS	23	6	513535	missing
6	CANTU RUZ ESCOTO MARIA JOSE	18	3	577715	missing
7	ESTAVILLO URREA JUAN PABLO	19	2	425199	missing
8	GARCIA VERDUZCO MARIO ABEL	19	12	286804	8370557679
9	GONZALEZ MANRIQUE PAULINA MILENKA	18	8	686234	3497507667
10	JIMENEZ OROZCO ANDREA	20	11	551318	7065060736
11	MUNGUIA QUINTERO AXEL FRANCISCO	19	9	524791	7142573985
12	NUÑO GUEVARA ALBERTO ENRIQUE	22	2	405879	6386313360
13	ORTEGA LARES SOPHIA	18	7	990031	9474160231
14	ORTIZ TIRADO GONZALEZ ESTEBAN	19	6	443880	8802712180
15	OSUNA RIOS RODRIGO	22	8	431962	8781791416
16	PARIS MERIN DOUGLAS FABIAN	21	4	770236	missing
17	REYES VALDEZ IRENE	21	6	330802	1020817685
18	RODRIGUEZ RAMIREZ FRANCISCO RICARDO	22	6	141546	7709903210
19	RUIZ PADILLA LUIS ANGEL	21	11	941912	missing
20	SALINAS GANDARA NOE ALEJANDRO	20	7	458383	1518977060
26	CASTILLO POMA ARACELI SOLEDAD	21	10	338719	4912740197
27	CONTRERAS GONZALEZ ANDREA LIZETTE	22	3	595136	7880944230
28	CONTRERAS TORRES YAQUIE GUADALUPE	19	11	382142	7729001070
29	CORTES SOTO ALEJANDRA	20	11	580550	6075198058
30	CRUZ DELGADILLO ANA CRISTINA	21	13	210028	9145396860
31	FLORES OROZCO OSCAR ALFONSO	20	1	529444	missing
32	GAMEZ ORTIZ LIDIA NATASHA	19	5	835054	7023063345
33	GARCIA GOMEZ ISAMAR	20	10	missing	missing
34	GUTIERREZ RAMIREZ EVELYN	21	8	113201	2752970583
35	LOMELI SALADO HANNIA CECILIA	20	7	347325	1980639143
36	MARTINEZ GOMEZ TAMARA VALERIA	21	1	562662	missing
37	MAYORQUIN HIGUERA ANEHI KARELY	23	5	899065	4796349262
38	NUÑO TISCAREÑO CARLOS ELIAS	22	9	427314	8563589265
39	PEÑA HINOJOSA HERMELA	20	12	missing	8979376292
40	PIEDRAS AYALA ANDRES	21	9	810821	5522227210
41	PINEDO TALANGO MARIA FERNANDA	20	9	missing	4902285262
42	PINTADO DELFIN MANUEL	20	8	394042	9820997302
43	RAMIREZ CAMBERO JOB	19	2	219378	7238529731
44	RAMIREZ DE LA ROSA LUIS MARIA	21	10	301775	4112594583
45	RAMIREZ HINOJOSA DIANA LAURA	20	10	444554	1384116632
46	SALCEDO CELIS PAULINA	22	8	925671	3453712953
47	VILLA DOMINGUEZ OSCAR GERMAN	19	2	569441	7367553692

La limpieza de datos es un proceso muy tardado debido a que se debe identificar “manualmente” los errores en la base de datos. Posterior a esto se pueden reparar algunos de los errores, sin embargo, otros requieren un poco más de tiempo. La limpieza de números telefónicos, así como de números de expediente es un proceso muy ágil, por contraparte nos encontramos con los nombres propios; en este apartado existen demasiados errores que no se limpian tan fácilmente (el principal problema es que hay algunos símbolos en dónde debería de haber espacios, si se eliminan directamente los símbolos quedan palabrasunidas a otras.)

5. Considerando la siguiente base de datos (en el Moodle se encuentra el archivo csv si lo quieren usar para la importarlo), responda las siguientes preguntas y justifique su respuesta:

nombre	2014	2013	2012	2011	2010
Casos de Dengue	1446	2584	560	175	1171
Casos de Influenza A H1N1	608	14	592	6	108
Casos de VIH/SIDA	506	630	978	665	578
Egresos hospitalarios	221364	155789	180462	220280	199288
Esperanza de vida al nacer	75.36	75.36	75.89	77.28	77.07
Muertes maternas	52	33	35	36	48

a. (0.5 punto) Realicé un gráfico donde muestre la evolución porcentual de cada una de las categorías al pasar de los años. Es decir, cual fue el cambio porcentual de cada una de las enfermedades de un año a otro.



Index	ıde:	0	1	2	3
Casos de De…	0	-0.850555	2.2	3.61429	-0.440402
Casos de In…	1	-0.944444	97.6667	-0.976351	42.4286
Casos de VI…	2	0.150519	0.470677	-0.355828	-0.196825
Egresos hos…	3	0.105335	-0.180761	-0.136721	0.420922
Esperanza d…	4	0.0027248	-0.0179865	-0.00698379	0
Muertes mat…	5	-0.25	-0.0277778	-0.0571429	0.575758

El cambio porcentual de 2011 a 2012 en casos de influenza H1N1 cambió dramáticamente (de 6 incidencias en 2011 a 592 en 2012)

b. (0.5 punto) Porcentualmente ¿qué categoría es la que ha tenido variaciones mayores en el transcurso de los años?

Los casos de H1N1 de 2011 a 2012 con cerca de un crecimiento del 97%, seguido de un decrecimiento del .97% y posteriormente una alza de 42%. Todos los demás se han comportado bastante tranquilos en relación a la incidencia de h1n1 (todos menores a 4%)

c. (0.5 punto) ¿Qué categorías han tenido un desempeño porcentual similar en el transcurso de los años?

Index		0	1	2	3	4	5
Casos de Dengue	0	0	4.45044	3.17989	3.01856	2.82257	2.55903
Casos de Influenza A H1N1	1	4.45044	0	4.15898	4.11234	4.0285	3.77814
Casos de VIH/SIDA	2	3.17989	4.15898	0	0.167492	0.371639	0.845683
Egresos hospitalarios	3	3.01856	4.11234	0.167492	0	0.226916	0.733182
Esperanza de vida al nacer	4	2.82257	4.0285	0.371639	0.226916	0	0.522232
Muertes maternas	5	2.55903	3.77814	0.845683	0.733182	0.522232	0

Se puede observar que las distancias euclidianas de los cambios porcentuales del dengue y de la influenza con respecto a las 4 variables restantes (todos normalizados) son muy altas, por lo tanto podemos inferir que los casos de VIH/SIDA, egresos hospitalarios, esperanza de vida al nacer y muertes maternas están muy relacionados entre sí, pero poco relacionados con los casos de influenza y del dengue.

(0.5 punto) ¿Qué años son los más parecidos considerando todas las categorías?

	0	1	2	3	4
0	0	2.25498	3.89444	3.77545	2.98691
1	2.25498	0	3.72127	4.55245	4.21945
2	3.89444	3.72127	0	4.05174	4.19568
3	3.77545	4.55245	4.05174	0	4.49178
4	2.98691	4.21945	4.19568	4.49178	0

normalizando todas categorías y aplicandoles un índice de similitud de distancia euclideana se puede encontrar que todos los años difieren mucho entre sí. Los más parecidos son el 1

y el 0 (2010 y 2011), seguidos del 0 con el 4 (2010 y 2014). No obstante, debido a estar normalizados, lo esperado era tener valores cercanos a 1 para poder considerarlos similares entre sí. En este caso parecen estar todos notoriamente diferentes.

Código:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.spatial.distance as sc

#%% Importar datos.
cancelacion = pd.read_csv('cancelacion_2017.csv',encoding='latin-1')
dirty1 = pd.read_csv('dirty_Info_Alumnos_v1.csv',encoding='latin-1')
dirty2 = pd.read_csv('dirty_Info_Alumnos_v2.csv',encoding='latin-1')
enfermedades = pd.read_csv('enfermedades.csv',index_col='nombre')

#%% Uso de suelo.
uso_suelo = pd.value_counts(cancelacion['Uso Suelo'])

plt.figure(figsize=(16,4))
uso_suelo.plot(kind='bar')
plt.show()

#%% Municipios con uso de suelo habitacional.
municipios = cancelacion[cancelacion['Uso Suelo'] == 'HABITACIONAL']
municipios_ = pd.value_counts(municipios['Municipio'])

plt.figure(figsize=(16,3.5))
municipios_.plot(kind='bar')
plt.show()

#%% Dejar únicamente datos con dígitos en teléfono.
clean2 = dirty2.copy()

clean2.iloc[:,-1] = clean2.iloc[:,-1].apply(only_digits) #para correr esta parte tienen que haber

```

```
#%% Eliminar números telefónicos con menos de 10 dígitos.
for i in range(len(clean2)):
    if len(str(clean2.iloc[i,-1])) != 10:
        clean2.iloc[i,-1] = 'missing'

#%% Eliminar expedientes incorrectos.
for i in range(len(clean2)):
    if len(str(clean2.iloc[i,3])) != 6:
        clean2.iloc[i,3] = 'missing'

#%% Limpiar semestre
clean2['sem estre'] = clean2['sem estre'].apply(only_digits)

#%% Limpiar nombres
clean2.iloc[:,0] = clean2.iloc[:,0].apply(replace,args=('¾','Ñ'))
esp = ['&','_','-','?','ð',':P',':0','x0',':)','%','/','.',':','     ','    ','   ','  ']
for i in esp:
    clean2.iloc[:,0] = clean2.iloc[:,0].apply(replace,args=(i,' '))
clean2.iloc[:,0] = clean2.iloc[:,0].apply(replace,args=('0','O'))
clean2.iloc[:,0] = clean2.iloc[:,0].apply(replace,args=('PEDRQ','PEDRO'))
clean2.iloc[:,0] = clean2.iloc[:,0].apply(remove_punctuation)
clean2.iloc[:,0] = clean2.iloc[:,0].apply(remove_digits)

#%% Enfermedades

#%% Cambiar orden, de izquierda a derecha.
enf = np.zeros((5,6))
for i in range(len(enf)):
    enf[i,:] = enfermedades.iloc[:,4-i]
enf = enf.T

#%% Graficar cambios porcentuales en las enfermedades.
pct = enf[:,1:]/enf[:,:-1]-1
pct = pd.DataFrame(pct)
pct.T.plot()

#%% Cambios porcentuales más parecidos.
norm = (pct-pct.mean(axis=0))/pct.std(axis=0)
dist = sc.squareform(sc.pdist(norm,'euclidean'))

#%%
enf = enf.T
enf1 = (enf-enf.mean(axis=0))/enf.std(axis=0)
anios = sc.squareform(sc.pdist(enf1,'euclidean'))

```

```
import string
#%% Funcion para retirar signos de puntuación.
def remove_punctuation(x):
    try:
        x = ''.join(ch for ch in x if ch not in string.puntuation)
    except:
        pass

    return(x)

#%% Remover digitos
def remove_digits(x):
    try:
        x = ''.join(ch for ch in x if ch not in string.digits)
    except:
        pass

    return(x)

#%% quitar espacios
def remove_whitespace(x):
    try:
        x = ''.join(x.split())
    except:
        pass

    return(x)

```

```
def replace(x,to_replace,replacement):
    try:
        x = x.replace(to_replace,replacement)
    except:
        pass

    return(x)

#%% convertir a mayusculas
def uppercase_text(x):
    try:
        x = x.upper()
    except:
        pass

    return (x)

#%%
def lowercase_text(x):
    try:
        x = x.lower()
    except:
        pass

    return(x)

#%%
def only_digits(x):
    try:
        x = ''.join(ch for ch in x if ch in string.digits)
    except:
        pass

    return(x)

```