UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
MAESTRÍA EN CIENCIA DE DATOS

APRENDIZAJE AUTOMÁTICO

Profesor: Jose Anastacio Hernandez Saldaña

Tarea 1: Análisis Exploratorio

Dora Alicia Guevara Villalpando
Matrícula: 1551003
Grupo: 003

30 de junio de 2024

UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Tarea 1. Análisis Exploratorio

Introducción.

El presente documento presenta los resultados obtenidos durante el proceso de realizar la Tarea 1 de Aprendizaje Automático. Esta tarea consiste en realizar el análisis exploratorio a una base de datos proporcionado en clase, esta base de datos cuenta con 5 columnas y 636,201 filas.

Desarrollo.

Durante la realización de esta tarea se llevaron a cabo cinco actividades principales:

1. Identificar las entidades más representativas.
2. Obtener estadísticas descriptiva de cada entidad.
3. Hacer agrupaciones por las entidades y sacar estadisticas de las agrupaciones.
4. Crear imagenes de estas estadisticas, ya sean histogramas, graficas de pastel, etc.
5. De alguna de las agrupaciones, hacer una pureba ANOVA.

A continuación se describen los hallazgos en cada una de las actividades realizadas.

Nuestra base de datos esta conformada por 5 columnas (variables), las cuales representan diferentes tipos de datos a analizar:

```
Data columns (total 5 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   Nombre       636201 non-null  object 
 1   Sueldo Neto  636201 non-null  float64
 2   dependencia  636201 non-null  object 
 3   Fecha        636201 non-null  object 
 4   Tipo         636201 non-null  object 
dtypes: float64(1), object(4)
```

En esta base de datos contamos con variable del tipo numérico (float): *Sueldo Neto*, y las demás estan clasificadas “objetos” de caracteres: *Nombre* (nombre del empleado), *dependencia* (nombre de la institución registrada en la UANL), *Fecha* y *Tipo* (clasificación de las dependencias registradas en administración, centro, preparatoria, facultad, hospital y otro).

Previo a iniciar con los análisis de estadística descriptiva decidí buscar si en las variables de la base de datos hay valores nulos para poder identificarlos y decidir que hacer con ellos; sin embargo, esta busqueda dio como resultado que en la base de datos no hay valores nulos por lo que procedí a obtener los estadísticos descriptivos.

Análisis descriptivos.

	Tipo	Cantidad_Empleados	Total_sueldo	Salario_Promedio	Salario_minimo	Salario_maximo	Desv_est
0	ADMIN	76994	1.024343e+09	13304.198963	177.20	147051.59	9923.529504
1	CENTRO	25117	3.211438e+08	12785.914565	245.38	120970.94	8274.192377
2	FACULTAD	274527	4.533969e+09	16515.567721	177.20	144501.40	11037.232652
3	HOSPITAL	105549	1.016583e+09	9631.383249	175.41	85007.77	4168.207964
4	OTRO	12190	1.391335e+08	11413.737791	285.73	78959.29	5779.618889
5	PREPARATORIA	141824	2.025400e+09	14281.078522	187.54	115258.74	8185.434724

Se analizaron los estadísticos descriptivos de las dependencias utilizando la categría pre-establecida en la columna *Tipo*. En estos resultados se observa que el sueldo acumulado (total_sueldo) máximo se encuentra en el Tipo = FACULTAD, así como el salario promedio más alto de 16,515.57.

Se decició obtener la desviación estándar de cada *Tipo* ya que nos sirve para hacer una estimación sobre cómo de dispersos están los datos con respecto a la media. Se observa que todas las desviaciones estándar son altas y esto representa mayor dispersión de los datos conforme a la media; esto se puede concluir al observar el rango tan amplio que hay entre el Salario mínimo y el Salario máximo.

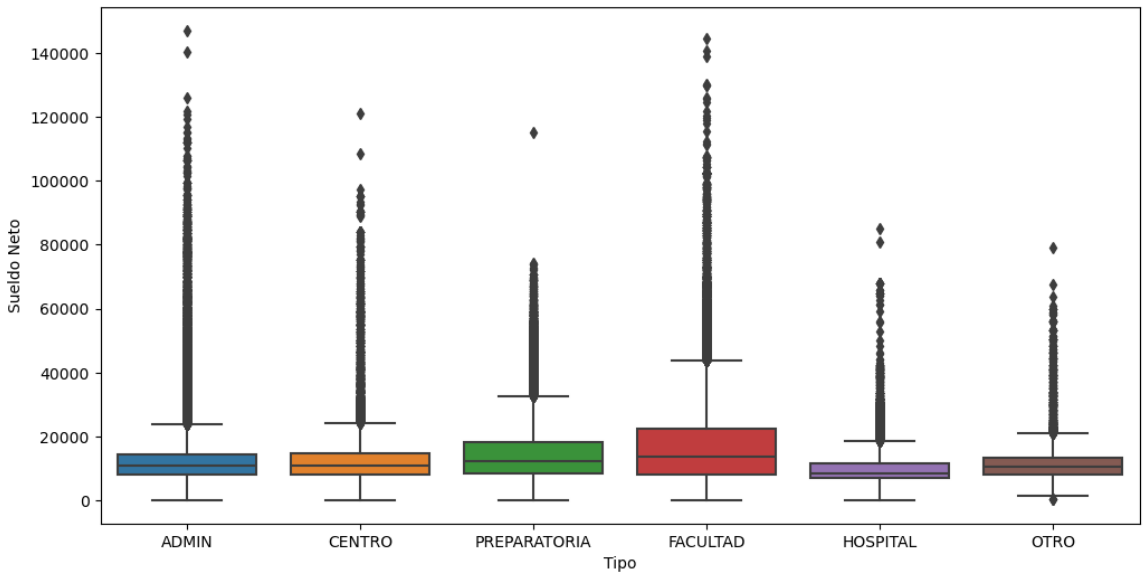
También se hizo un análisis similar al de la tabla anterior pero agregando el año, esto para identificar patrones o diferencias. Se observó que en 2019 se tiene una cantidad menor de empleados registrados y por ende un sueldo acumulado menor a los demás años. La excepción sería la información que se muestra del 2024, ya que al ser el año corriendo la información no está completa.

Con las estadísticas descriptivas se identifica que la FACULTAD durante el 2023 es donde se tiene la mayor cantidad de empleados y se maneja el salario máximo. Se observó lo siguiente:

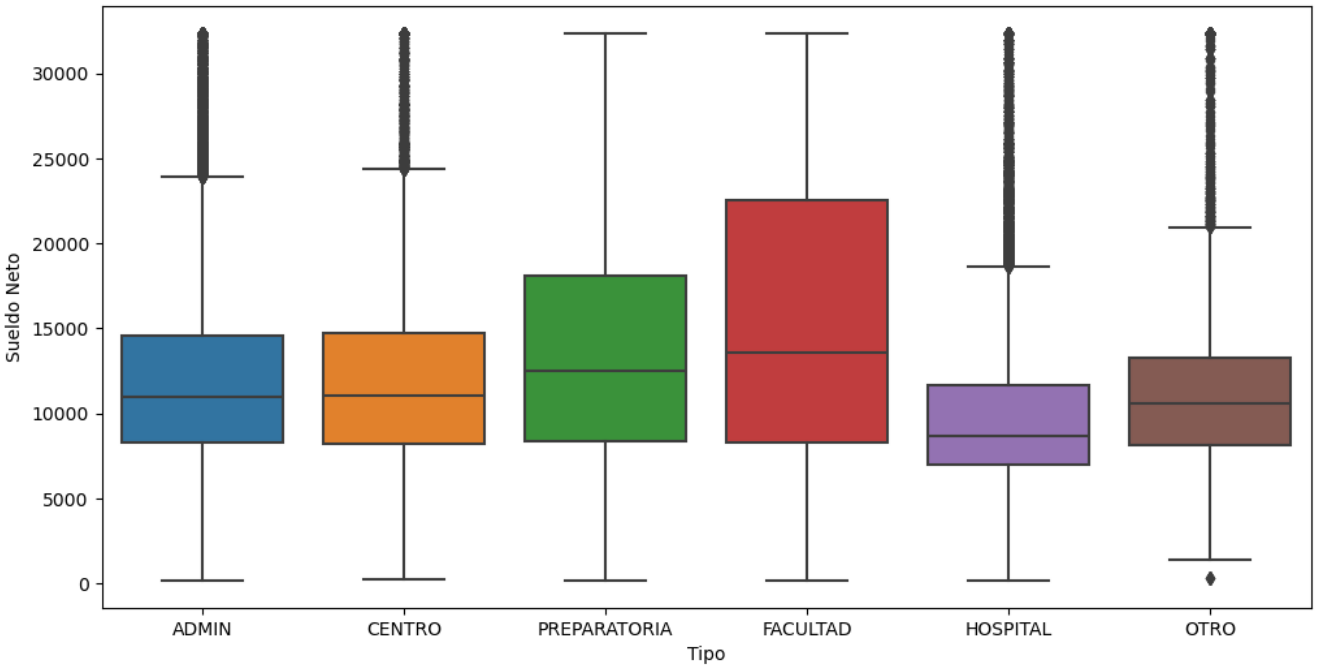
- La facultad que cuenta con el salario promedio más alto es *FIME*.
- La facultad que cuenta con el salario máximo es *Ciencias Biologicas*.
- La facultad de *Artes Visuales* cuenta con el salario promedio más bajo de todas.
- La facultad de *Artes Escenicas* cuenta con el salario máximo más pequeño de todas las facultades aunque su salario promedio es mayor al de la facultad de artes visuales.

dependencia	Total Sueldo Neto	Salario Promedio	Sueldo Mínimo	Sueldo Máximo
FAC. DE AGRONOMIA	26,136,915.30	19,190.10	608.85	125,886.32
FAC. DE ARQUITECTURA	60,091,246.79	15,294.29	608.85	79,814.57
FAC. DE ARTES ESCENICAS	9,409,731.16	14,748.79	1,305.09	59,995.95
FAC. DE ARTES VISUALES	19,078,718.18	14,698.55	2,591.59	68,648.00
FAC. DE CIENCIAS BIOLOGICAS	78,184,644.05	20,312.98	622.32	144,501.40
FAC. DE CIENCIAS DE LA COMUNICACION	33,213,386.94	16,825.42	637.65	60,812.93
FAC. DE CIENCIAS DE LA TIERRA	16,688,414.76	21,758.04	1,102.41	79,219.82
FAC. DE CIENCIAS FISICO-MATEMATICAS	54,217,220.15	18,422.43	608.85	69,929.26
FAC. DE CIENCIAS FORESTALES	19,148,200.45	18,411.73	1,217.70	85,062.92
FAC. DE CIENCIAS QUIMICAS	71,357,821.57	18,376.98	425.10	61,470.14
FAC. DE CONTADURIA PUBLICA Y ADMON.	88,788,370.79	16,181.59	775.39	90,982.69
FAC. DE ECONOMIA	15,061,717.35	20,632.49	4,106.18	79,018.64
FAC. DE ENFERMERIA	23,603,062.81	17,483.75	2,218.16	64,993.90
FAC. DE FILOSOFIA Y LETRAS	48,974,236.74	17,111.89	405.90	129,748.17
FAC. DE ING. CIVIL	35,094,853.05	17,373.69	811.80	101,727.07
FAC. DE ING. MECANICA Y ELECTRICA	202,850,267.58	23,267.98	405.90	118,943.89
FAC. DE MED. VETERINARIA Y ZOOT.	22,269,281.01	17,064.58	348.02	70,765.45
FAC. DE MEDICINA	178,060,919.19	21,079.78	474.68	129,748.17
FAC. DE MUSICA	14,330,168.17	15,036.90	608.85	69,288.06
FAC. DE ODONTOLOGIA	36,530,450.56	19,277.28	1,934.38	86,275.97
FAC. DE ORGANIZACION DEPORTIVA	33,098,204.13	16,137.59	405.90	111,296.88
FAC. DE PSICOLOGIA	35,693,611.53	16,964.64	405.90	80,386.20
FACULTAD DE CIENCIAS POL√çTICAS Y RELACIONES INTERNACIONALES	38,518,141.98	17,050.97	608.85	78,185.81
FACULTAD DE DERECHO Y CRIMINOLOGIA	63,909,907.69	16,484.37	608.85	67,797.71
FACULTAD DE SALUD PUBLICA Y NUTRICION	25,445,670.30	15,678.17	608.85	78,631.96
FACULTAD DE TRABAJO SOCIAL Y DESARROLLO HUMANO	22,455,783.81	18,391.31	608.85	69,673.32

Se realiza una gráfica de cajas para cada uno de los *Tipos* de dependencias:



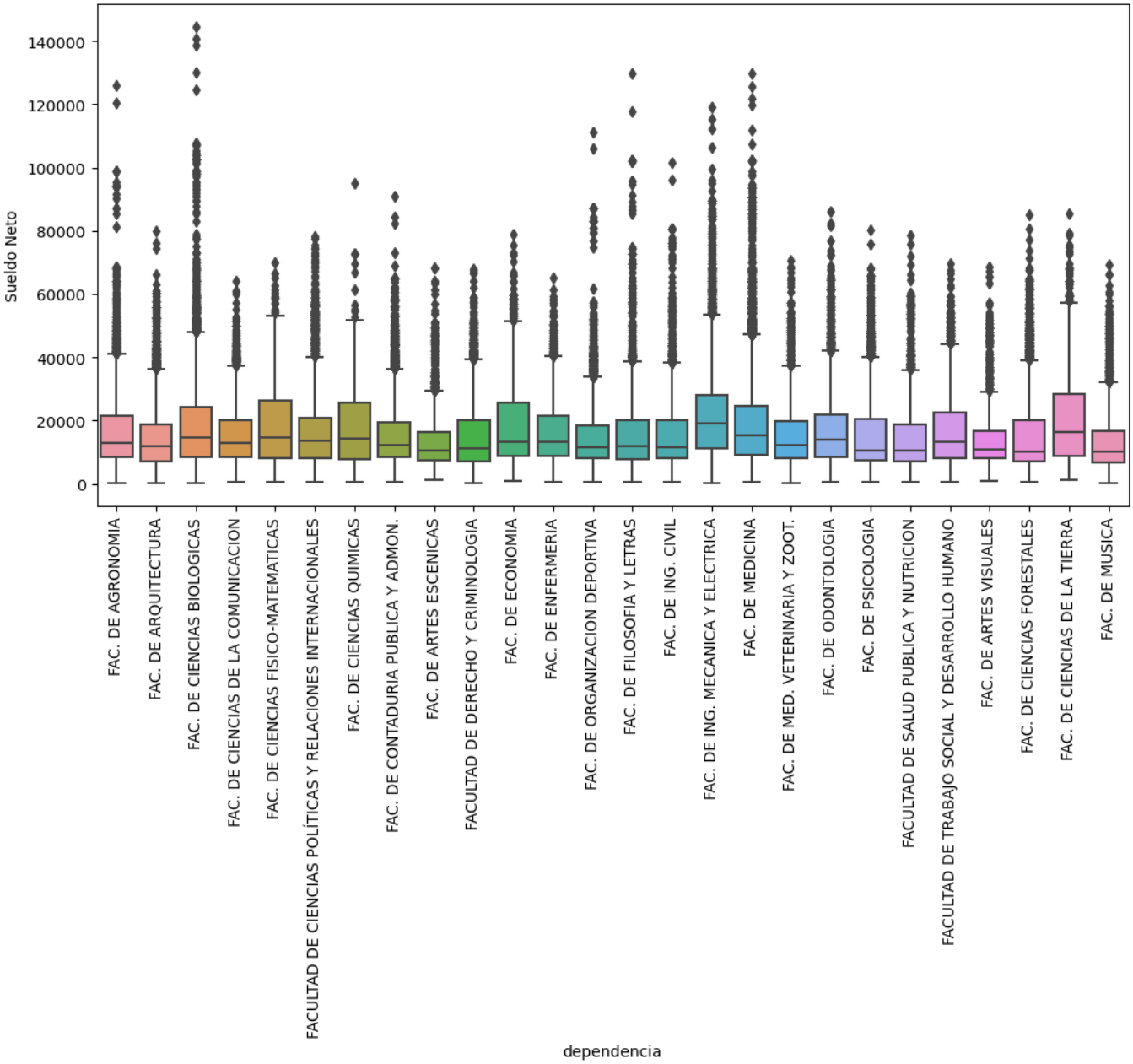
Reducimos los outliers para poder apreciar el diagrama de caja de una mejor forma:



Se concluye lo siguiente:

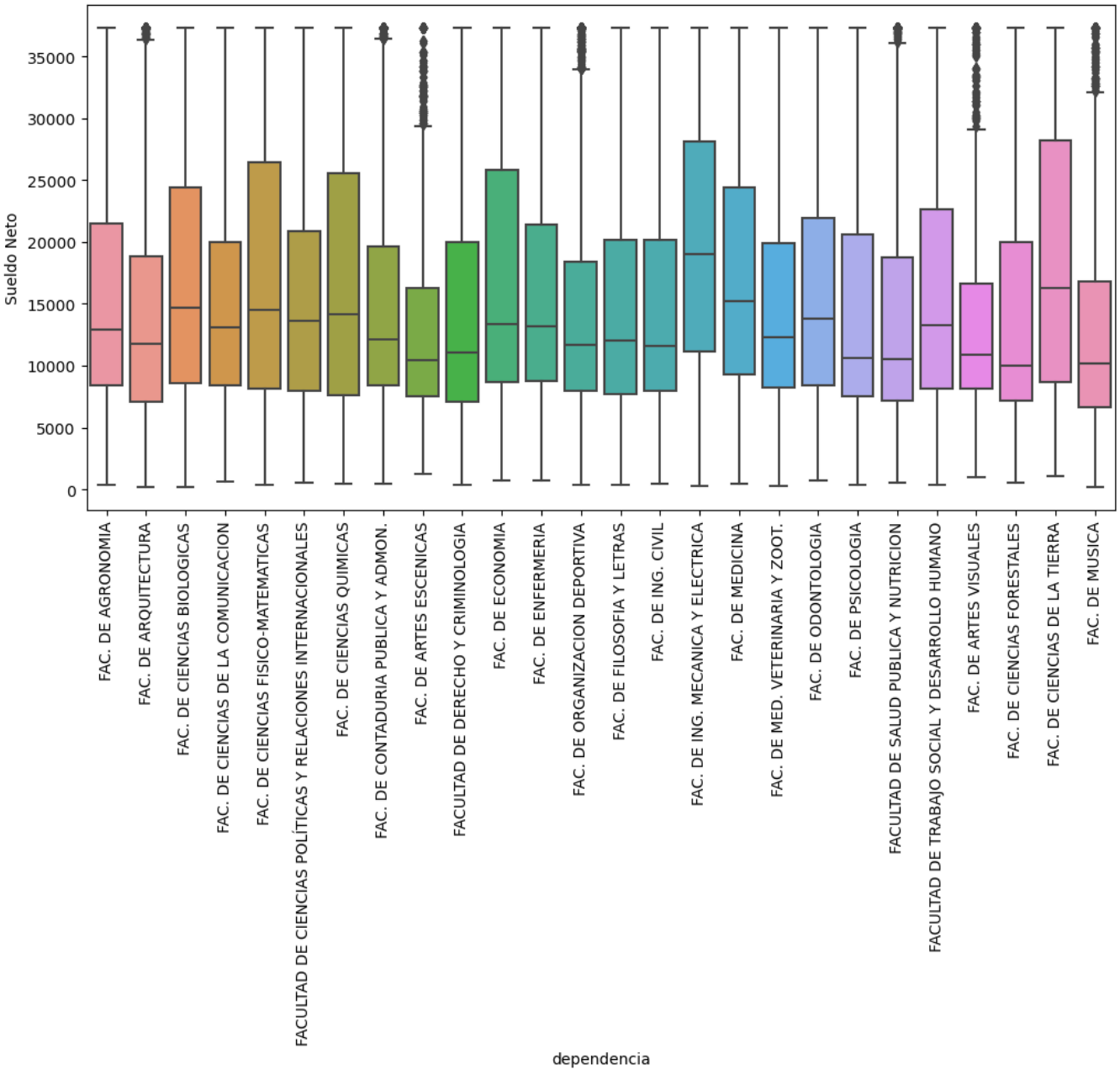
- La caja de FACULTAD es más amplia que las demás.
- Todas tienen valores atípicos (outliers).
- La caja que parece mostrar menor dispersión en los datos alrededor de la media es OTRO.
- Se observa que el salario promedio de ADMIN y CENTRO es similar.
- El salario promedio más alto se encuentra en FACULTAD,
- El salario promedio más pequeño se encuentra en HOSPITAL.

A continuación se presenta un conjunto de gráficos de caja para cada una de las facultades:



La facultad de *ciencias biologicas* cuentan con un rango más disperso de *sueldo neto*.

Reducimos outliers para observar mejor la información:



- FIME cuenta con el salario promedio más alto.
- La segunda facultad con el salario promedio más alto es MEDICINA.

ANOVA:

$H0$: No hay diferencias entre las medias de los diferentes grupos.

$H1$: Al menos un par de medias son significativamente distintas la una de la otra.

```
Estadístico de prueba: 16.27327327327326
Valor p: 0.006105667507561778
Se encontraron diferencias significativas entre los grupos.
```

Como $p - valor = 0.0061$ entonces rechazamos $H0$.

Se infiere que al menos un par de medias son signifivativas por lo que no hay normalidad.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
ADMIN	CENTRO	-108956761.73	0.9654	-523422711.4793	305509188.0193	False
ADMIN	FACULTAD	559280012.1017	0.0036	144814062.3524	973745961.8509	True
ADMIN	HOSPITAL	8678739.4183	1.0	-405787210.3309	423144689.1676	False
ADMIN	OTRO	-137631537.2517	0.9109	-552097487.0009	276834412.4976	False
ADMIN	PREPARATORIA	169700558.0467	0.8111	-244765391.7026	584166507.7959	False
CENTRO	FACULTAD	668236773.8317	0.0004	253770824.0824	1082702723.5809	True
CENTRO	HOSPITAL	117635501.1483	0.9523	-296830448.6009	532101450.8976	False
CENTRO	OTRO	-28674775.5217	0.9999	-443140725.2709	385791174.2276	False
CENTRO	PREPARATORIA	278657319.7767	0.3422	-135808629.9726	693123269.5259	False
FACULTAD	HOSPITAL	-550601272.6833	0.0042	-965067222.4326	-136135322.9341	True
FACULTAD	OTRO	-696911549.3533	0.0002	-1111377499.1026	-282445599.6041	True
FACULTAD	PREPARATORIA	-389579454.055	0.075	-804045403.8043	24886495.6943	False
HOSPITAL	OTRO	-146310276.67	0.8879	-560776226.4193	268155673.0793	False
HOSPITAL	PREPARATORIA	161021818.6283	0.842	-253444131.1209	575487768.3776	False
OTRO	PREPARATORIA	307332095.2983	0.2435	-107133854.4509	721798045.0476	False