UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

PRACTICA 1: ANÁLISIS EXPLORATORIO.

APRENDIZAJE AUTOMATICO

CÉSAR ALEJANDRO HERNÁNDEZ OROZCO

Matricula: 1990010

GRUPO 003

MAESTRO: JOSÉ ANASTASIO HERNÁNDEZ SALDAÑA

INTRODUCCIÓN:

En esta practica usaremos el conjunto de datos proporcionado por la Universidad Autonoma de Nuevo Leon para realizar un analisis exploratorio de estos datos y aprender cosas de ellos. Con este analisis vamos a explorar, resumir y visualizar un conjunto de datos para extraer información útil y detectar patrones, tendencias, relaciones o anomalías que puedan estar presentes en los datos

Este analisis tendra como objetivo aprender a manejar mejor los datos usando la librería Pandas de Python y buscamos profundizar en los datos contenidos dentro de este dataset.

A traves de este estudio, se busca identificar patrones, tendencias y posibles disparidades en los ingresos reportados por cada entidad dentro de la universidad. En este analisis se llevarán a cabo agrupaciones según las entidades identificadas para obtener estadísticas agregadas y comparativas. A través de la visualización gráfica, utilizando técnicas como histogramas, gráficos de pastel y otros, se presentarán las tendencias y distribuciones observadas en los datos. Se enfatizará la creación de múltiples imágenes que faciliten la comprensión visual de los hallazgos obtenidos durante el análisis exploratorio.

DESARROLLO:

1.- Identificar las entidades (que es lo que puede ser sujeto de estudio) del conjunto.

- Sueldo neto: Representa el salario neto que recibe cada empleado despues de deducciones y reteneciones. Es uno de los datos mas utilizados para la realización de este reporte y nos ayuda a hacer muchas comparaciones.
- Dependencias: Indica la unidad organizativa o dependencia dentro de la UANL a la cual pertenece cada empleado. Nos ayudara a agrupar mejor los datos.
- Fecha: Puede referirse a la fecha en que se realizó el pago del sueldo o la fecha de registro de los datos. Nos ayudara a realizar analisis temporal.
- Tipo de empleado: Clasificación que indica la categoría o tipo de empleado dependiendo de a que parte de la UANL esta aliado.

Estas son las entidades identificadas de este conjunto de datos. A continuación podremos ver como se desarrollo este analisis.

2.- Obtener las estadisticas (descriptiva) de cada entidad (min, max, avg, std)

Usando las herramientas proporcionadas por Pandas obtenemos una descripción de los datos.

	Sueldo_Neto	Fecha
count	636201.000000	636201
mean	14241.682401	2021-12-01 04:03:39.394813952
min	175.410000	2019-12-01 00:00:00
25%	8007.660000	2020-11-01 00:00:00
50%	11426.500000	2021-11-01 00:00:00
75%	17654.630000	2023-01-01 00:00:00
max	147051.590000	2024-01-01 00:00:00
std	9578.442311	NaN

Con esta tabla podemos empezar a obtener información bastante relevante de nuestro conjunto de datos. De este tabla podemos ver que:

- Nuestro conjunto de datos cuenta con 636201 registros. Este fue un dato que me sorprendio ya que no esperaba que alla tanta gente registrada en la nomina de UANL incluso contando todas las instituciones.
- La persona con mayor salario llega a recibir 147,051 pesos.
- La persona con menor salario solo recibe 175.241 pesos.
- La mayoria de los empleados tienen un salario que ronda los 14,241 pesos.

Con estos datos obtenidos se plantean varias preguntas y reflexiones. Es especialmente interesante indagar en las diferencias significativas entre el salario más alto y el más bajo, lo que nos lleva a explorar qué tipo de roles y responsabilidades tienen estas personas y cómo se justifica esta disparidad salarial dentro de la institución.

Para conocer un poco mas se uso el codigo para obtener mas detalles sobre los empleados con los salarios extremos:

El empleado con el menor salario es LUCY ELENA FRANCISCO BAUTISTA que pertenece a HOSPITAL UNIVERSITARIO con 175.41

El empleado con el mayor salario es SANTOS GUZMAN LOPEZ que pertenece a RECTORIA con 147051.59

Estos ejemplos específicos nos permiten comprender mejor las dinámicas salariales dentro de la UANL y generar nuevas preguntas sobre equidad salarial, roles dentro de la universidad y políticas de compensación.

3.- Hacer agrupaciones por las entidades y sacar estadisticas de las agrupaciones.

Lo siguiente que haremos en este reporte es realizar agrupaciones usando las entidades del conjunto de datos para explorar patrones y caracteristicas especiales dentro de las distintas dependencias de la UANL. Al agrupar los datos por entidades podemos obtener estadísticas descriptivas que nos ayuden a entender mejor la distribución y las tendencias dentro de cada grupo.

Empezamos realizando agrupaciones usando las dependencias usando su sueldo neto. Con ello podemos obtener algunas conclusiones con respecto a las dependencias:

Dependencia	count	mean	std	min	25%	50%	75%	max
HOSPITAL UNIVERSITARIO	105549.0	9631.383249	4168.207964	175.41	6946.200	8629.98	11622.83	85007.77
FAC. DE ING. MECANICA Y ELECTRICA	35035.0	20479.725818	11402.530515	297.42	11146.905	19061.10	28131.80	118943.89
FAC. DE MEDICINA	34295.0	18330.532933	11872.347237	474.68	9261.540	15221.28	24443.29	129748.17
ESC.IND.Y PREPA.TEC.ALVARO OBREGON	21477.0	13136.464283	7824.823355	205.36	7767.260	10788.57	17009.20	70605.09
FAC. DE CONTADURIA PUBLICA Y ADMON.	20089.0	14751.867493	8678.563533	474.28	8392.600	12085.65	19621.03	90982.69

El conjunto de datos completo puede ser encontrado dentro del repositorio.

Usando estas herramientas cada vez empezamos a conocer mas de como operan los distintos departamenteo de la universidad. Con estas tablas podemos ver que con mucha diferencia el Hospital universitario es la dependencia con mayor cantidad de empleados en su nomina.

Despues para conocer un poco mas de los datos dados por estos reportes entonces tambien queremos saber quienes son las personas que mas ganan y menos ganan de cada dependencia:

Empleados con mas salario en cada dependencia:

```
Empleados con el salario más alto por dependencia:
                                    Nombre  Sueldo_Neto  \
0              HECTOR LUIS AGUILAR GONZALEZ     68323.61
1              RAFAEL AMADOR MAYORGA OLVERA     35590.67
2                      CESAR MORADO MACIAS     43505.89
3                SOCORRO GUAJARDO GONZALEZ     77130.15
4    JOSE JAVIER VILLARREAL ALVAREZ TOSTADO     78959.29
..                                     ...          ...
147    CLAUDIA DE MONSERRAT SAMANIEGO GARZA     67532.31
148              MARTHA NORA ALVAREZ GARZA     87384.51
149            SERGIO MANUEL SANCHEZ TREJO     49809.89
150              EUGENIO JOSE REYES GUZMAN     58697.17
151          MARIA DE LA PAZ BELTRAN ORTEGA     35182.17

                                      Dependencia
0                    AUDITORIA INTERNA DE LA U.A.N.L.
1    C. INNOVACION; INVEST. Y DESLLO. DE INGENIERIA...
2                          C.DE ESTUDIOS HUMANISTICOS
3                       C.DE INV.Y DES.DE ED.BILINGUE
4           CAPILLA ALFONSINA BIBLIOTECA UNIVERSITARIA
..                                                 ...
147                            TEATRO UNIVERSITARIO
148                               TESORERIA GENERAL
149                           UNIDAD DE TRANSPARENCIA
150                  WORLD TRADE CENTER MONTEREY-UANL
151       ´CAPILLA ALFONSINA´ BIBLIOTECA UNIVERSITARIA
```

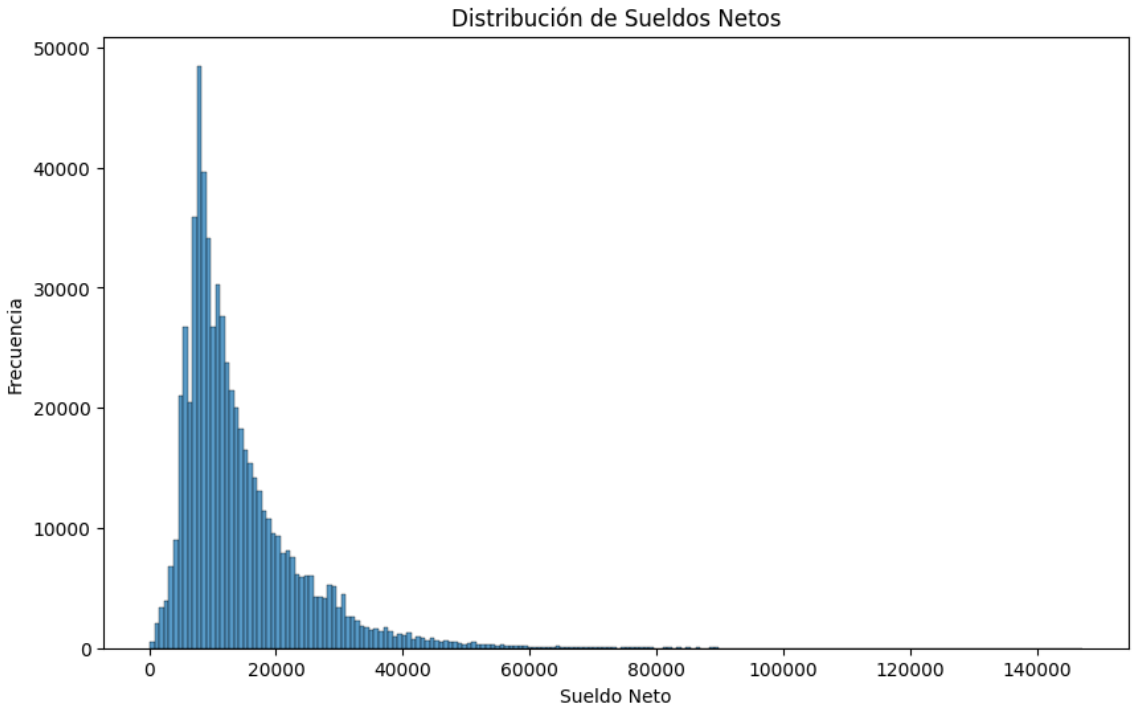
Empleados con menos salario en cada dependencia:

```
Empleados con el salario más bajo por dependencia:
                                Nombre  Sueldo_Neto  \
0    SAMANTHA ELIZABETH GUERRERO SALAZAR      2694.80
1        JESUS EDUARDO SANDOVAL MARTINEZ      5045.30
2          ROSA CECILIA JUAREZ RODRIGUEZ      2895.11
3             MARCY NALLELY AVILA GUZMAN      2903.41
4          NAYELLI CRISTINA VELEZ GARCIA      1535.48
..                                   ...          ...
147                IDALIA CASTRO GARDUÑO      2611.44
148        BRENDA LIZETT ELIZONDO ARENAS       245.38
149          IVONNE JANETH IXBA SANTIAGO       285.73
150           OLIVIA KARYNA VEGA ALCAZAR      7184.08
151              LUIS FIDEL CAMACHO PEREZ       311.22

                                          Dependencia
0                    AUDITORIA INTERNA DE LA U.A.N.L.
1    C. INNOVACION; INVEST. Y DESLLO. DE INGENIERIA...
2                          C.DE ESTUDIOS HUMANISTICOS
3                       C.DE INV.Y DES.DE ED.BILINGUE
4           CAPILLA ALFONSINA BIBLIOTECA UNIVERSITARIA
..                                                 ...
147                            TEATRO UNIVERSITARIO
148                               TESORERIA GENERAL
149                           UNIDAD DE TRANSPARENCIA
150                  WORLD TRADE CENTER MONTEREY-UANL
151       ´CAPILLA ALFONSINA´ BIBLIOTECA UNIVERSITARIA

[152 rows x 3 columns]
```

Usando estos datos y tambien los de las anteriores conclusiones podemos empezar a conocer cuales son las dependencias que tienen a sus empleados con mejores salarios.

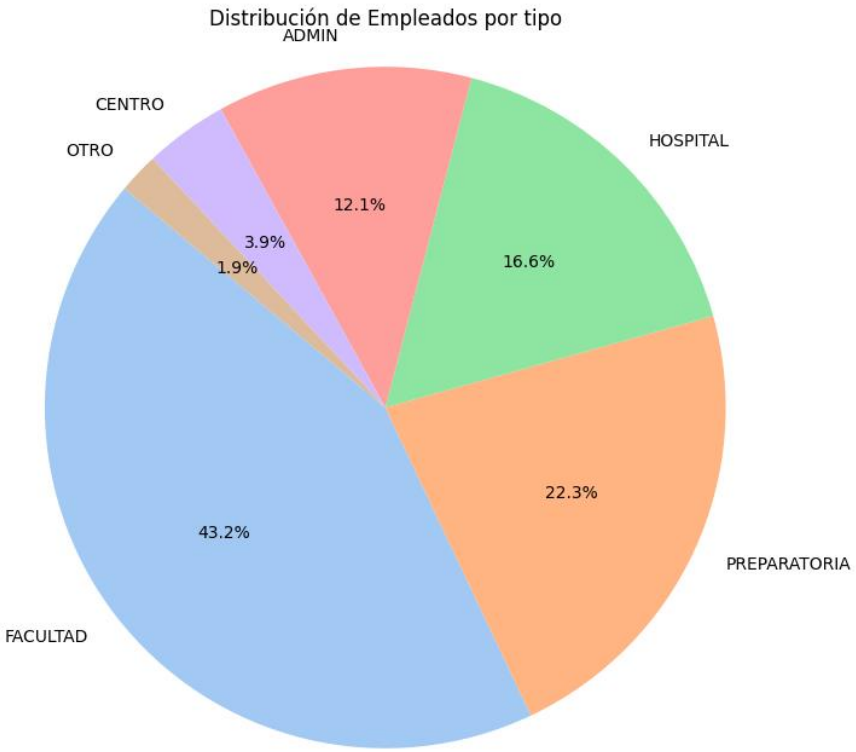
4. Crear imagenes de estas estadisticas, ya sean histogramas, graficas de pastel, etc, al menos 3 graficas diferentes, Hay que crear muchas imagenes, asi que creen ciclos y recorran las agrupaciones.

Ya con los datos obtenidos anteriormente empezamos a conocer un poco mas sobre como se distribuyen los distintos datos de los salarios. Para empezar a sacar mejores conclusiones empezaremos a generar las graficas para poder ver mejor los datos. Empezaremos en ver la distribución de los sueldos netos.



Con lo visto anteriormente podemos ver que tal y como se veia en la media de los salarios la mayoria de los empleados tienen un salario cercano a los 15000 pesos y de ahí podemos ver como poco a poco va bajando y vemos que la gran mayoria cuenta con un salario por debajo de los 50000 pesos y escasean mucho las demas.

Ahora, sabemos que los empleados tienen distintos tipos dependiendo de en que area de la UANL trabajen. Por medio de una grafica de pastel podemos ver como se dividen los distintos empleados por tipo:

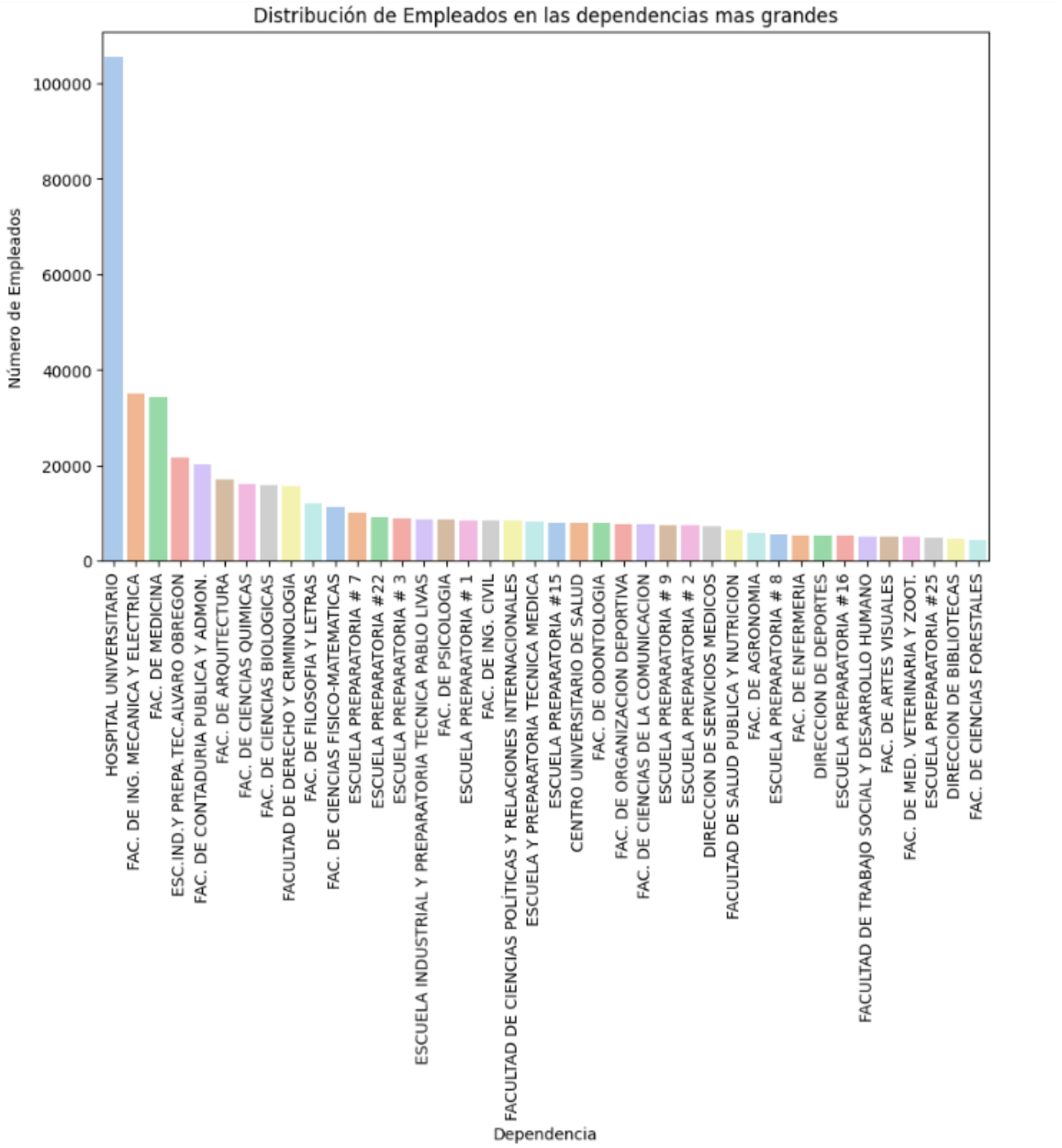


Esta grafica es bastante interesante debido a que nos muestra de que manera se distribuyen los empleados por sus tipos:

1. La facultad tiene mas empleados con 43.2% empleados.
2. Con un 22.3% sigue la preparatorio en empleados.
3. El hospital cuenta con 16.6% de empleados.
4. La universidad cuenta con el 12.1% de sus empleados trabajando en administracion.
5. 3.9% trabajan en el centro de la UANL.
6. 1.9% tienen algun otro tipo de puesto.

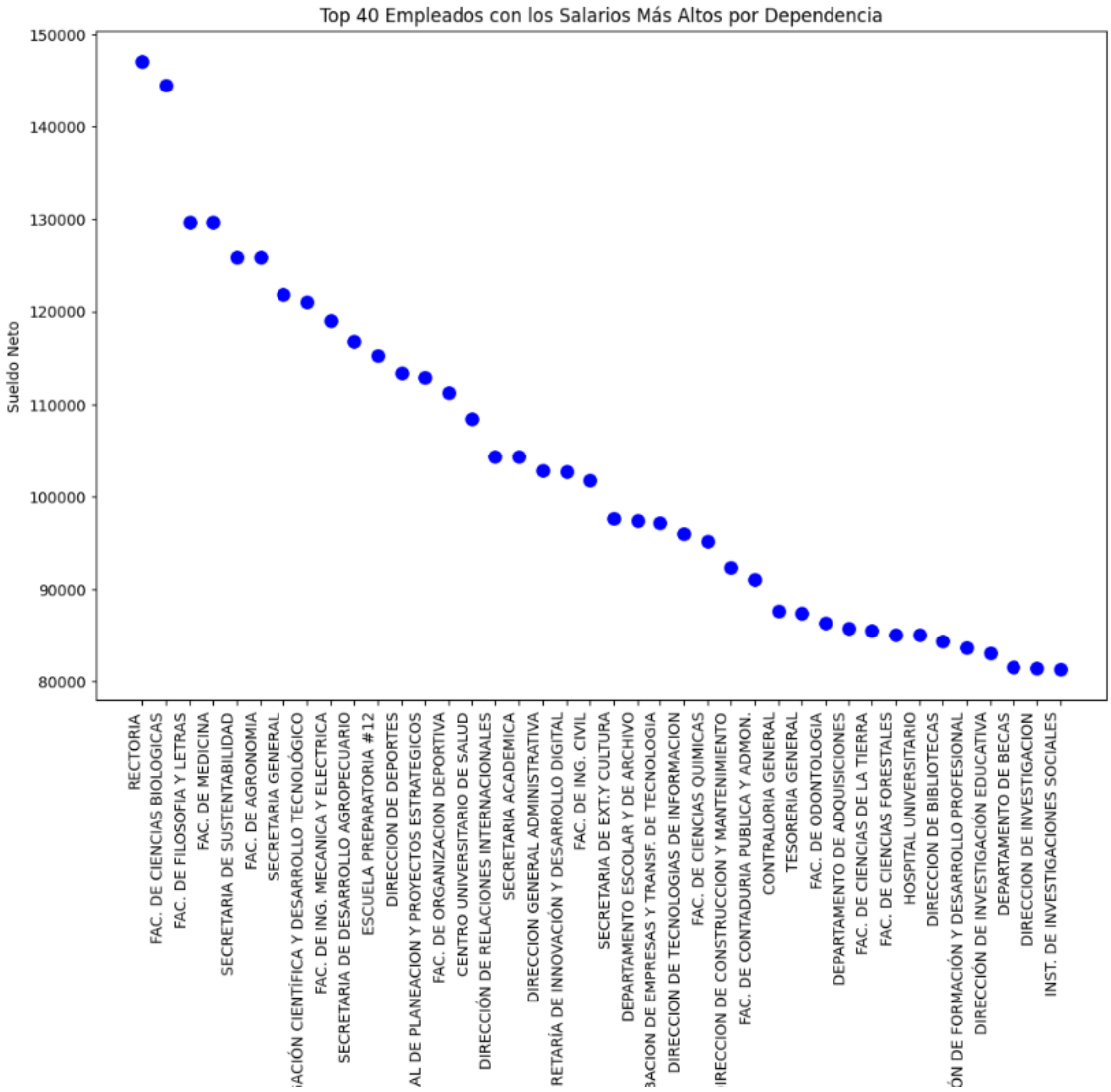
Con estos datos podemos confirmar que la universidad sigue siendo la mayor fuente de empleos dentro de la UANL.

Anteriormente hemos empezado a obtener datos de las dependencias. Ahora estamos interesados en conocer como se comparan los numeros de empleados entre ellas. Debido a que son demasiadas dependencias decidimos graficar las 20 facultades con mas empleados.

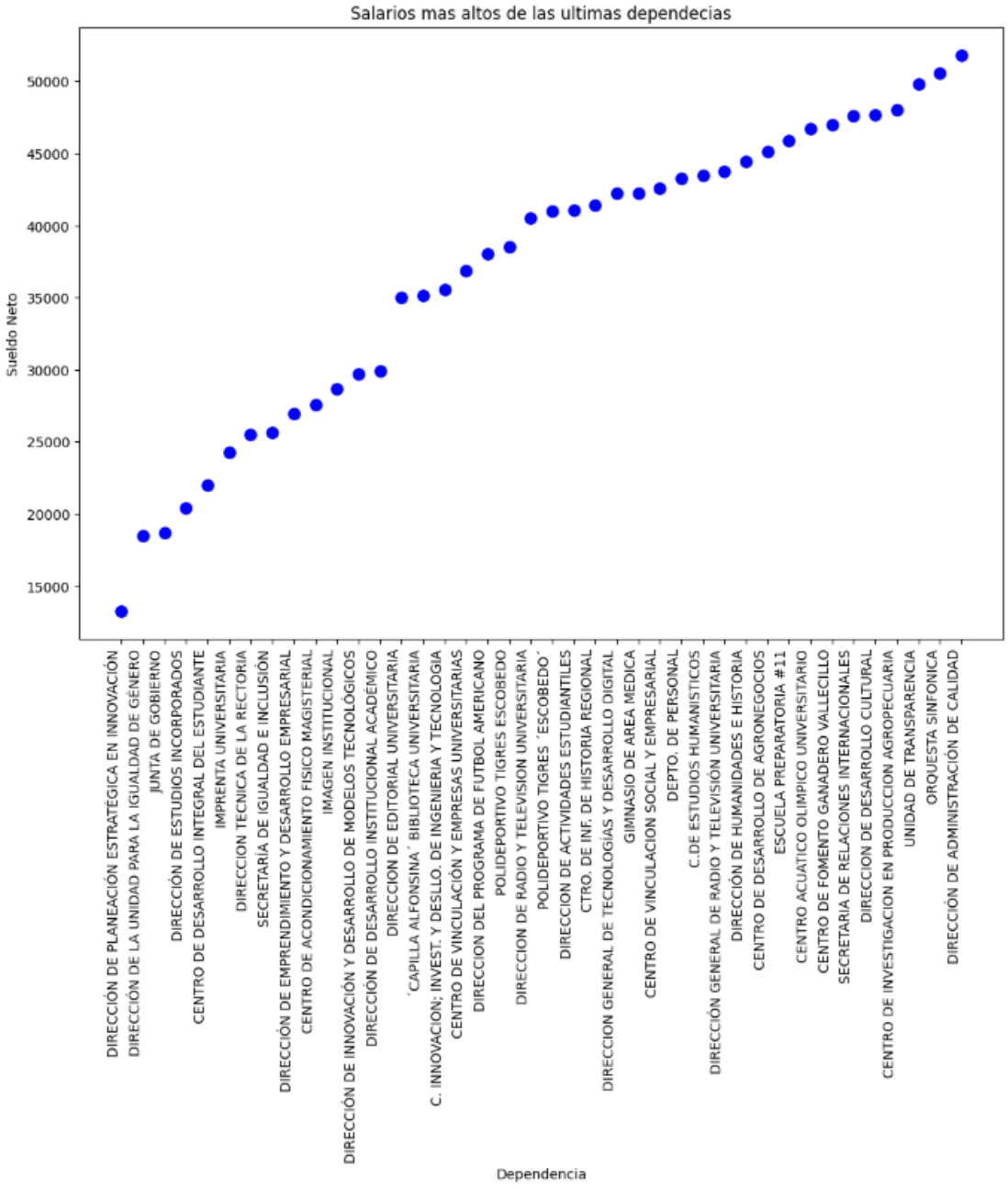


Es interesante porque nos muestra que el hospital universitario es la dependencia con mayor numero de empleados con mucha diferencia. Este dato lo podiamos empezar a ver anteriormente con los datos ya obtenidos pero ahora vemos que esta diferencia es bastante grande. Cuenta con mas del doble de empleados que el 2do lugar la facultad de Ingenieria Mecanica y Electrica.

Ya conociendo como se van distribuyendo los empleados en las distintas dependencias nos interesa conocer otros datos usando las dependencias. Anteriormente ya obtuvimos quienes eran los empleados con mayor salario dentro de cada dependencia, ahora queremos empezar a verlo graficamente comparando los empleados con mayor salario de cada dependencia:

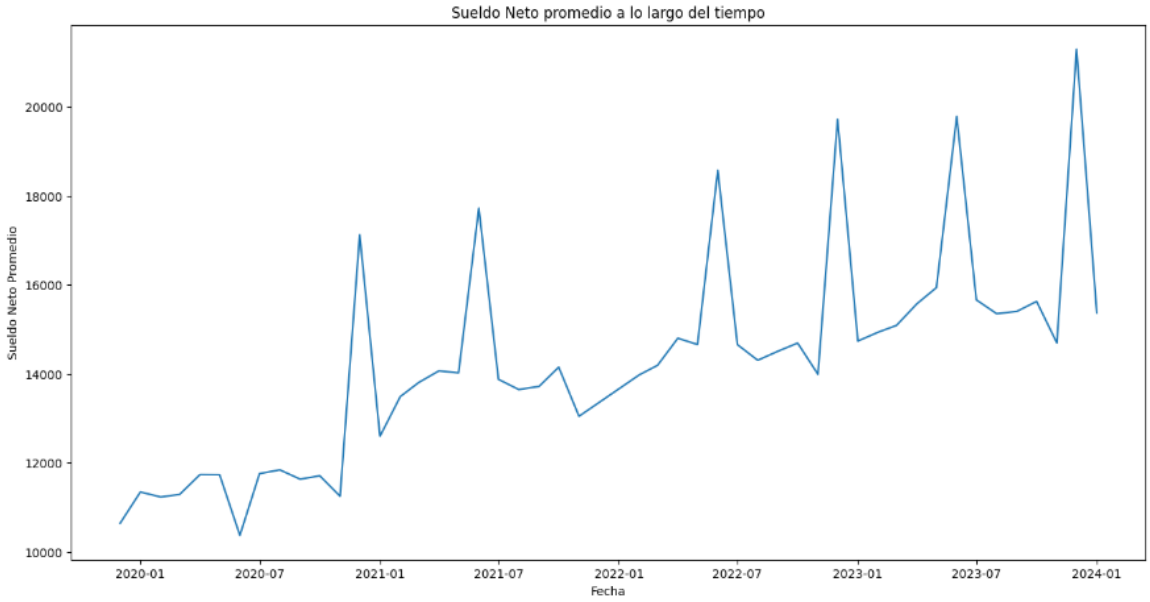


Por medio de esta grafica podemos empezar a conocer que al menos en el area de los empleados con mayor salario no hay una diferencia tan grande entre el mayor salario de cada dependencia, pero solo dentro de las que reciben mas dinero. Ahora, como seria el salario mas alto de las dependencias con menores salarios mas altos:



Es interesante esta grafica porque nos muestra que existen dependecias donde el salario mas alto no pasa ni siquiera de los 20,000 pesos en contraste con otras dependecias que llegan incluso a recibir mas de 100,000 pesos sus empleos.

Ahora, hablando de sueldos seria interesante conocer como se han ido adaptando estos al paso del tiempo, ya sea para combatir la inflación o para dar mejor prestaciones a sus empleados:



Ahora, podemos ver que desde el 2020 que es donde inician nuestros datos los salarios han subido de una manera bastante estable, con picos de vez en cuando, estos probablemente pueden deberse a algun evento como ingreso de empleados o tambien puede deberse a cuando ingresan los datos a la base de datos.

Durante la realización de este reporte se realizaron tambien graficas para mostrar la distribucion de los sueldos netos de cada facultad. Estas imágenes pueden encontrarse en los archivos adjuntos en este repositorio.

5.- De alguna de las agrupaciones, hacer una pureba ANOVA, para ver si hay diferencias entre los elementos. recordando, hay que probar que las muestras son o no normales. si son normales anova para la prueba y t-student para saber quien es el diferente. Si la muestra no es normal, prueba con kruskall-wallis y tukey para saber quien es el diferente.

Agruparemos los datos por dependencia y empezaremos a realizar prueba de normalidad a cada una:

```
JUNTA DE GOBIERNO: stat = 1.0, p-value = 0.0
RECTORIA: stat = 1.0, p-value = 0.0
SECRETARIA GENERAL: stat = 1.0, p-value = 0.0
DIRECCION GENERAL DE PLANEACION Y PROYECTOS ESTRATEGICOS: stat = 1.0, p-value = 0.0
SRIA. DE INVESTIGACIÓN CIENTÍFICA Y DESARROLLO TECNOLÓGICO: stat = 1.0, p-value = 0.0
CENTRO DE VINCULACION SOCIAL Y EMPRESARIAL: stat = 1.0, p-value = 0.0
SECRETARIA DE SUSTENTABILIDAD: stat = 1.0, p-value = 0.0
SECRETARIA DE DESARROLLO AGROPECUARIO: stat = 1.0, p-value = 0.0
DIRECCION GENERAL ADMINISTRATIVA: stat = 1.0, p-value = 0.0
CONTRALORIA GENERAL: stat = 1.0, p-value = 0.0
AUDITORIA INTERNA DE LA U.A.N.L.: stat = 1.0, p-value = 0.0
TESORERIA GENERAL: stat = 1.0, p-value = 0.0
OFICINA DE LA ABOGACÍA GENERAL: stat = 1.0, p-value = 0.0
DIRECCION DE PREVENCION Y PROTECCION UNIVERSITARIA: stat = 1.0, p-value = 0.0
DEPARTAMENTO ESCOLAR Y DE ARCHIVO: stat = 1.0, p-value = 0.0
DIRECCION DE RECURSOS HUMANOS Y NOMINAS: stat = 1.0, p-value = 0.0
DIRECCION DE TECNOLOGIAS DE INFORMACION: stat = 1.0, p-value = 0.0
```

Se realiza la prueba de Kolmogorow-Smirnow en cada una de las dependencias para comprobar la normalidad de la distribucion de los sueldos usando de hiposis nula que las muestras son normales.

Despues de realizar la prueba podemos ver que en todas las dependencias se obtiene un resultado de valor estadistico de 1.0 y un valor p de 0.0. Esto nos dice que se rechaza la hipotesis en todas las dependencias. Estos nos dice que ninguna de las dependencias tiene una distribución normal con sus salarios.

Debido a que ninguna de las distribuciones son normales podemos recurrir a una prueba no parametrica para determinar si existen diferencias significativas entre las medianas de mas de dos grupos independientes. La prueba seleccionada para compararlas es la prueba de Tukey.

Usando las herramientas proporcionadas por Python podemos comparar todas las dependencias entre ellas y sacar conclusiones de cada una:

group1	group2	meandiff	p-adj	lower	upper	reject
AUDITORIA INTERNA DE LA U.A.N.L.	C. INNOVACION; INVEST. Y DESLLO. DE INGENIERIA Y TECNOLOGIA	-539.0509	1.0	-2812.0213	1733.9196	False
AUDITORIA INTERNA DE LA U.A.N.L.	C.DE ESTUDIOS HUMANISTICOS	2598.3148	0.0985	-109.8845	5306.5141	False
AUDITORIA INTERNA DE LA U.A.N.L.	C.DE INV.Y DES.DE ED.BILINGUE	1509.5787	0.0	383.3225	2635.835	True
AUDITORIA INTERNA DE LA U.A.N.L.	CAPILLA ALFONSINA BIBLIOTECA UNIVERSITARIA	-1204.4825	0.0521	-2411.8391	2.8741	False

Aquí encontramos algunas de las conclusiones obtenidas al comparar las dependencias entre ellas. Podemos ver que al menos entre estos datos existe una diferencia significativa entre las medias de la Auditoria interna de la UANL y el Centro de investifación y desarrollo de educación bilingüe.

Podemos ver muchas mas de estas comparaciones en el archivo tukey_results.csv adjuntado a este reporte.

Conclusión:

Toda la realización de este proyecto nos permitio conocer mucho sobre como operan los pagos dentro de la UANL. Es interesante ver de manera grafica la gran diferencia que existen entre los salarios dentro de la institución. Me sorprendio ver como mientras las gran mayoria cuenta con un sueldo que ronda los 15,000 pesos existe gente que llego a ganar mas de 100,000 y no son pocos. Dentro de la universidad existe una gran variabilidad salarial y nos plantea muchas dudas de a que se debera esto. Desconocia que la institución con mas empleados dentro de la institución es el hospital universitario, es interesante conocer lo importante que es para la universidad.

Este reporte me ayudo mucho a mejorar mis habilidades para programar usando Python y obtener información interesante de un conjunto de datos. Es interesante ver la gran diferencia de salarios que existe incluso dentro de una misma dependencia.

En conclusión, este analisis nos ayuda a tener una base solidad para futuras investigaciones que se realicen sobre la UANL, la realización de este reporte me desperto la curiosidad de conocer un poco mas sobre como se determina el salario que tiene una persona dentro de la institución.