Gente querida, buenas tardes!!

En nuestro último encuentro trabajamos un tema particular, que si bien no se relacionaba con el 1er TP que han estado desarrollando, me pareció súper importante: la revisión preliminar del archivo de datos, y en un caso como el propuesto, el arreglo de los mismos.

Trabajamos diferentes herramientas para conseguir esto: ¡recuerden que Excel puede ser de gran ayuda en la visualización y exploración inicial de los datos!

**Lists-Sets-Dicts Comprehentions**

Apareció nuevamente, en el código, el uso de corchetes para armar "***listas por comprensión***". Sobre eso les recomendé que revisen el material disponible de las materias en las que ya vieron ese contenido, además de algunas de las muchísimas Webs donde encontrarán explicaciones y ejemplos a montones. A modo de ayudita, aquí les recomiendo estas:

* Básico, cortito y al pié  
  <https://www.w3schools.com/python/python_lists_comprehension.asp>
* Más profundo, arranca desde mucho más atrás y llega más lejos  
  <https://www.datacamp.com/community/tutorials/python-list-comprehension>

Recomiendo incluso que vean el detalle de agregar condicionales a una generación de conjuntos de datos con List Comprehentions o Set Comprehentions o Dict Comprehentions

**Group-By: Split / Apply / Combine**

También conversamos brevemente sobre esto de las agrupaciones por un campo clave (o varios) e incluso con la posibilidad de calcular funciones de agregado tales como media, desvío, máximo, mínimo, etc...

Recomiendo ponerse en claro con este tema ya que lo necesitaremos para los próximos trabajos.

Les dejo un par de links:

* <https://realpython.com/pandas-groupby>  
  Ver nota de presentación y código desarrollado como soporte para la lectura + pruebas e investigación
* <https://pandas.pydata.org/docs/user_guide/groupby.html>  
  Referencia oficial Pandas GroupBy: Split, Apply, Combine

También les dejo algunos recursos útiles sobre los que estuvimos conversando en las últimas clases, y que serán de utilidad en este TP y los venideros:

* **Generación de Histogramas** organizando o agrupando por intervalos de clase las variables que tenemos. Sobre esto les recomiendo este artículo: <https://www.codigopiton.com/como-hacer-un-histograma-en-python/>
* **Scatter y Scatter 3D**: nos permitirá observar la distribución de valores que tenemos e incluso observar a simple vista los rangos normales y tendencias de una variable en función de las otras.

**Plotly**

Hemos recomendado en varias oportunidades esta biblioteca. Para las gráficas antes indicadas, y para muchas otras -especialmente gráficas interactivas-, ¡la recomiendo totalmente!

Encontrarán una guía de referencia muy completa y ejemplificada en este [link: https://ichi.pro/es/visualizacion-con-plotly-express-guia-completa-258792796523351](https://ichi.pro/es/visualizacion-con-plotly-express-guia-completa-258792796523351)

Observarán que el post está lleno de publicidad... pero salvando esa pequeña incomodidad, la calidad de las gráficas que presenta es fabulosa, todo con su correspondiente código fuente, y con  buen nivel de explicación sobre los parámetros involucrados en cada método y sus funcionalidades!!

También les pido investigar/revisar particularmente, 2 temas:

* **Gráficas "descriptivas" de distribución de frecuencias**: los "**Box-Plots**" son de gran utilidad para presentar una versión “simplificada” de los histogramas clásicos de distribución de frecuencias: es fundamental entender la naturaleza de este tipo de gráficos. Y me permito agregar además las "**Gráficas de Violín**". Estas últimas resultan muy útil cuando además de la distribución queremos contrastar dos grupos bien definidos, quizá por una variable binaria o una clasificación de dos grupos, como por ejemplo, datos Masculinos y Femeninos. Resulta más útil incluso que los BoxPlots porque nos deja ver la forma de la distribución, no sólo sus parámetros estadísticos, y de manera espejada para ambas poblaciones. Les recomiendo la siguiente web donde está cubierto en detalle la anatomía de los BoxPlots y el significado de cada una de sus partes: <https://visualizationcheatsheets.github.io/boxplot.html>  
  *Deberemos recordar algunos elementos de estadística descriptiva como: mediana, cuartiles, y rango intercuartílico, para poder entender su significado  
    
  Podemos ver aquí una comparativa gráfica entre un histograma clásico de distribución de frecuencias y su representación abreviada con un Box Plot:*Gráfico, Histograma

  Descripción generada automáticamente

*En la gráfica siguiente puede verse una comparativa entre los Violin Plots y los Box Plots.*

Gráfico

Descripción generada automáticamente

* **Grado de correlación entre variables en estudio**: un recurso muy útil para establecer la independencia de las variables es la Matriz de Correlación. Pueden obtenerla de manera directa con pandas, e incluso graficarla de manera muy sencilla. Así, podrán ver cómo algunas variables son más independientes de las restantes que otras. El coeficiente de correlación para cada variable se presenta en distribución matricial, en una matriz simétrica, que nos permite ver claramente esto. Pueden investigar el método **.corr** de Pandas y otras fuentes, incluyendo **Plotly**, **Seaborn**, o **MatPlotLib** para representarlo en gradientes de colores.

Tabla

Descripción generada automáticamente

*Las variables se distribuyen por igual en filas  y columnas, y en cada intersección podemos observar su****coeficiente de correlación****.*

**Contenido específico TP#2**

Les dejo también en el adjunto el CSV para el siguiente TP, cuyas indicaciones siguen a continuación:

En este nuevo conjunto de datos deberán desarrollar la tarea de Análisis Exploratorio siguiendo los tips con que hemos estado trabajando, procurando sobre todo entender la naturaleza de los datos, los rangos y distribución de valores de cada variable, y eventualmente la interrelación que hubiere entre ellas.

Se trata de información sobre 5000 personas relativa a 5 variables: Altura, Peso, Edad, Sexo, y Presión Arterial Máxima.

El trabajo es, como lo anticipamos, esencialmente una tarea de EDA (Análisis Exploratorio de Datos), prestando especial atención a los siguientes puntos:

* Observar el DataSet con todas las herramientas posibles, familiarizarse con su contenido, con la naturaleza y rango de valores "válidos" para cada variable. Observar mínimos, máximos, mediana y desvío o varianza.
* Listo
* En función de lo anterior, determinar errores, omisiones, y outliers. Corregir el Dataset eliminando del mismo los registros inconsistentes: este es un lujo que podemos darnos ya que se trata de muchas mediciones!
* Listo
* Procurar establecer gráficas descriptivas que presenten la info contenida en el Dataset desde diferentes puntos de vista: la idea es comunicar todos los detalles posibles sobre esta población de manera sintética y gráfica.
* Listo
* Llegamos así a lo más importante de todo: Tenemos como OBJETIVO establecer el grado de correlación entre las variables; y más específicamente, en qué medida depende la PA del resto de las variables. Esto puede ser de manera directa o modificada. Quizá alguno encuentra que la PA no depende fuertemente de ninguna de las variables, pero sí de una combinación de ellas. Sugerencia: tener en cuenta algunas medidas antropométricas que pueden derivarse de los datos provistos, como por ejemplo, el IMC (índice de masa corporal).
* Listo
* También se puede, considerando lo anterior, establecer una suerte de regresión lineal para las variables que guarden cierto grado de interdependencia: sería interesante presentar los resultados obtenidos.
* Listo
* Tener en cuenta que el experimento debe ser "repetible", y por tanto, debe quedar registro documental de cada tarea realizada.
* En lo posible, parametrizar cada acción de modo tal que uno pueda ensayar el impacto ante diferentes modificaciones en dichos parámetros (en cada nueva corrida del experimento, obviamente). Por caso, un ejemplo de esto podría ser establecer cuántos desvíos consideraré como válidos a cada lado de la mediana, o de la media, de la distribución de valores de cada variable... Obviamente no es el único posible: quedará en el análisis que cada uno realice considerar qué más puede parametrizarse.

Deseo que disfruten de la tarea, que todos aprendamos mucho de la experiencia realizada, y obviamente estoy atento a los comentarios y seguimiento que vayamos realizando en el grupo de WhatsApp hasta nuestro próximo encuentro.