**UNIVERSIDAD SANTO TOMÁS. PARTE AAAA.**

**PARCIAL ESTADÍSTICA EXPLORATORIA**

**Docente: Edna Carolina Moreno**

**2022-2**

**Context**

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. Here's how the Hass Avocado Board describes the data on their website:

La tabla de datos representa registros semanales del volumen (en unidades) y precio en el mercado minorista de comercio de Aguacate. Los datos de escaneo minorista provienen directamente de las cajas registradoras de los minoristas basadas en las ventas minoristas reales de aguacates Hass. A partir de 2013, la siguiente tabla refleja un conjunto de datos de venta minorista expandido y de ventanilla múltiple. El precio promedio (de aguacates) en la tabla refleja un costo por unidad (por aguacate), incluso cuando las unidades múltiples (aguacates) se venden en bolsas. Los códigos de búsqueda de productos (PLU) en la tabla son solo para los aguacates Hass. Otras variedades de aguacates (por ejemplo, pieles verdes) no se incluyen en esta tabla.

Algunas columnas relevantes en el conjunto de datos son:

Date – Fecha de la observación

AveragePrice – Precio promedio por unidad

type - conventional or organic

year - the year

Region - the city or region of the observation

Total Volume - Total number of avocados sold

4046 - Total number of avocados with PLU 4046 sold

4225 - Total number of avocados with PLU 4225 sold

4770 - Total number of avocados with PLU 4770 sold

Acknowledgements

Many thanks to the Hass Avocado Board for sharing this data!

(Verifique cuidadosamente cuáles son las diferencias en estos tres archivos)

#1. Lea adecuadamente las bases de datos. Vea el siguiente ejemplo y lea los datos para el 2015, etc

de los años 2015 hasta el año 2018.

#2. (0.5) Elabore un Data Frame donde compare los nombres de las bases de datos del 2015 hasta el 2018.

#3. (1.0) Arregle los nombres de las columnas si es necesario. Una todas las bases de datos en un solo Data Frame: Qué precauciones debe tener?

Elimine las filas repetidas con la función unique.

#4. (1.0) Con la funciónn subset seleccione los datos correspondientes a las regiones de "NewYork", "Boston", "Orlando" y con esta nueva base de datos elabore una tabla donde compare las medidas descriptivas

de la variable total\_volume en términos del type (organico o convencional) INTERPERTE ESTA TABLA.

Nota: De aquí en adelante trabaje con esta base de datos filtrada es decir con los datos de las regiones de "NewYork", "Boston", "Orlando"

#5. (1.0)Con los datos del punto anterior (es decir con los registros correpondientes a "NewYork", "Boston", "Orlando")Elabore e INTERPRETE una tabla cruzada con la función tabyl donde se observe el porcentaje de registros de acuerdo a la Ciudad (region) y a tipo de augacate (Type**) asegúrese de que salga porcentaje del total columnas**.

INTERPRETE ESTA TABLA RESALTE LAS CARACETRÍSTICAS MÁS IMPORTANTES.

#6. (0.5) Cree una nueva columna en la base de datos donde indique el porcentaje de aguacates vendidos con PLU 4046 (4046 - Total number of avocados with PLU 4046 sold) en relación al total (Total Volume - Total number of avocados sold)

#7.(1.0) Use la función groupby and summarise para crear alguna tabla interesante. INTERPRETE