Entrega final proyecto

Samuel Fernando De Dios Pérez

Robert Daniel Fonseca Lesmez

Lukas Morera Torres

Diagrama entidad-relación

Esta es la versión final del diagrama, el cual ya se encuentra normalizado y con todas las relaciones en su debido orden:

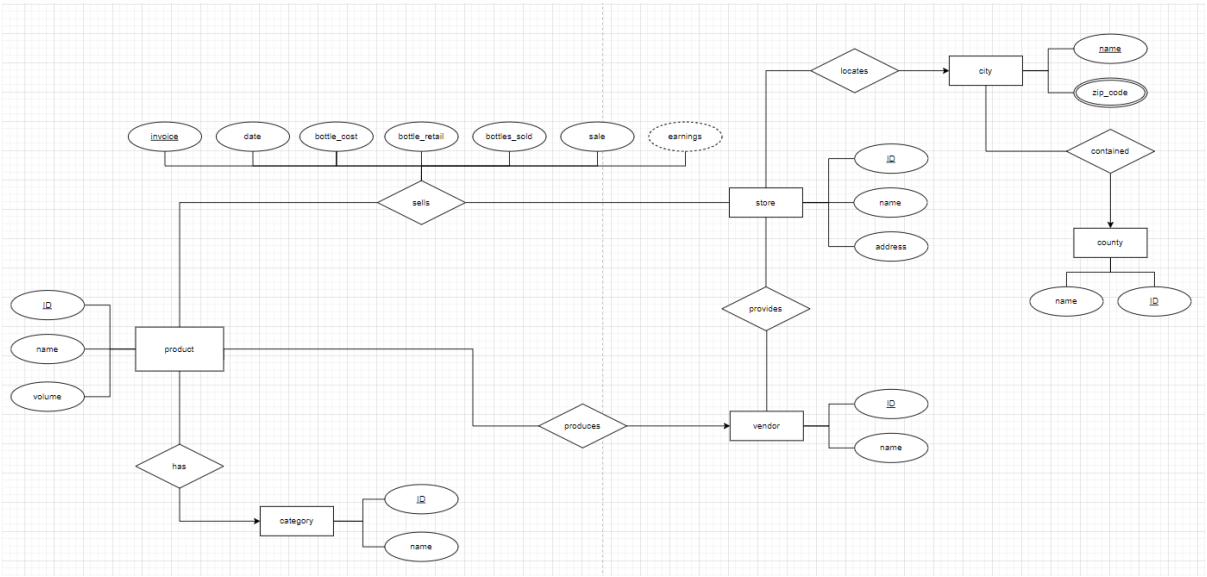
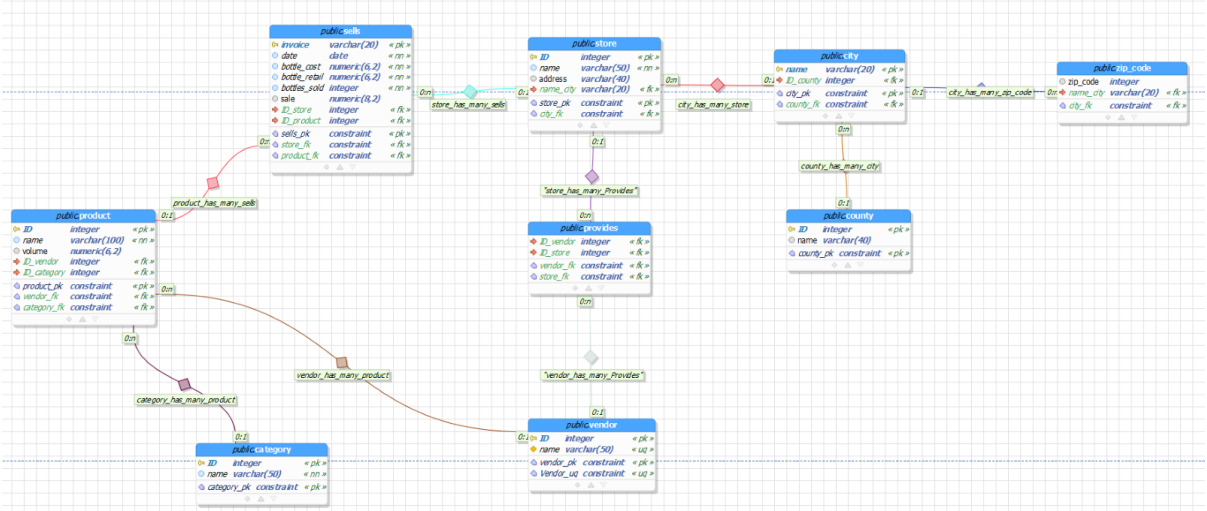


Diagrama relacional

Esta es la versión final del diagrama, donde ya se implementó el proceso de normalización, creando todas las relaciones correspondientes:



Una vez verificados los diagramas, procedemos a realizar las descripciones de las gráficas de los escenarios de análisis previamente definidos, realizados mediante Dash en Python:

Descripción de las situaciones de análisis

- **Ventas por categoría:** Se planteó el uso de 3 diagramas: barras, embudo y dispersión, esto al ser los que más claridad presentan con la cantidad de datos que estamos manejando. Los diagramas describen la cantidad de ventas según cada categoría de licor, mostrando la cantidad registrada en la base de cada una.
- **Ingresos anuales:** Se plateo el uso de 4 diagramas: barras, pie, línea y embudo, esto ya que son menos datos y se pueden hacer más representaciones de los mismos. En los diagramas se describe los ingresos que dejaron cada año las ventas de licor registradas en la base.
- **Ganancias mensuales por condado:** Se plantea el uso de 4 diagramas: barras, dispersión, calor y polar, esto ya que, nuevamente, son bastantes datos, y estos diagramas permiten una mejor visualización de los mismos, además de contar con la posibilidad de segmentar los datos, permitiendo una visualización más completa. En los diagramas se describen las ganancias o utilidades que dejaron las ventas de licor en cada condado de Iowa, nuestro lugar de estudio, separándolas por meses.
- **Ventas mensuales:** Se plantea el uso de 4 diagramas: barras, pie, línea y embudo, esto ya que son pocos datos, y son posibles más representaciones de los mismos. En los diagramas se describe la cantidad de ventas de licor por cada mes, esto según las ventas registradas en la base.
- **Costo por categoría:** Se plantea el uso de 3 diagramas: barras, dispersión y embudo, esto ya que son bastantes datos, y estos diagramas presentan una mejor visualización de los mismos. Las gráficas describen el costo total que representa en las tiendas cada categoría de licor registrada en la base.
- **Ingresos anuales por categoría de Vodka:** Se plantea el uso de 5 diagramas: barras, pie, dispersión, calor y polar, esto ya que son pocos datos, y se pueden visualizar de más maneras, además de poder segmentarlos para tener una visualización más completa de los datos. Los diagramas representan los ingresos por cada categoría de Vodka que esté registrada en la base, separándolos por cada año de la base.

Discusión

1. Ventas por categoría:

Samuel: De acuerdo con los diagramas, se puede evidenciar que la categoría que más tiene registrada ventas es “VODKA 80 PROOF”, con un total de 131.167 ventas, seguido por “CANADIAN WHISKIES” con 98.621 ventas. Estas dos categorías tienen ventas bastante mayores en comparación al resto de categorías, esto ya que la tercera más vendida, es decir “STRAIGHT BOURBON WHISKIES” cuenta con 57.974 ventas, siendo un numero qué, además de alejarse bastante de las 2 primeras categorías, es un valor similar al que tienen otras categorías, siendo posible evidenciar que no se cuentan con valores muy destacables en el resto de los datos, teniendo algunos un número de ventas que es casi imperceptible en la visualización de los datos.

Lukas: Se puede evidenciar que son bastantes las categorías que no gozan de muchas ventas, lo cual genera la sospecha de estas corresponden a tipos de licor muy específicos, y que, debido a su exclusividad o desconocimiento popular, no suelen tener mucha demanda en el mercado. Por otro lado, los más demandados se sospecha son tipos de licor más comunes y populares entre las personas, por tanto, son los que tienen más demanda en el mercado, ya que estas categorías tienen más productos en su haber. De esto se concluye que, si se quiere que un licor sea popular, debe ser de una categoría bien conocida entre los consumidores.

Daniel: Respecto a las gráficas, la de barras muestra un mayor número de datos, permitiendo ver de forma clara la frecuencia de cada una de las categorías; sin embargo, al ser bastantes datos, los mismos pueden verse algo apretados entre sí. La de embudo es similar a la de barras, permitiendo visualizar varios datos, con el agregado de que se puede apreciar mejor la comparativa entre categorías, pero de igual manera la gráfica se puede ver algo apretada dado el gran número de datos. Por último, la de dispersión presenta un mayor orden entre los datos, no apreciándose tan juntos entre sí; sin embargo, pueden presentarse algunas dificultades al ver que valor tiene cada punto, esto dado que el punto no es tan exacto en dar valores como los anteriores diagramas.

2. Ingresos anuales:

Samuel: De acuerdo con los diagramas, se puede ver que el año que más tuvo ingresos mediante las ventas de licor fue 2014, seguido de 2013, luego 2012 y por último 2015. Sin embargo, se puede apreciar que la diferencia de ingresos entre años no es tan grande, viéndose en diagramas como el de pie que tienen cantidades de ventas bastante similares entre sí. Por ello, se puede concluir que la venta de licor se mantiene constante a través de los años, teniendo unos ingresos entre 32M y 34M, que son los valores que se evidencian en estos años.

Lukas: Se puede apreciar que en todos los años se consiguen ingresos bastante similares, siendo en promedio 33M de dólares por año gracias a la venta de licor. El hecho de que no varie mucho esta cantidad con el paso de los años y de que genere bastantes ganancias, es un indicativo de que el mercado del licor es uno bastante rentable en Iowa, ya que genera bastantes ingresos y estos se mantienen estables con cada año que pasa, generando más seguridad para invertir en este mercado.

Daniel: Respecto a las gráficas, la de barras permite ver los datos de manera clara al ser pocos datos; sin embargo, dada la poca diferencia entre los datos puede ser algo difícil la comparativa. Algo similar ocurre con el de embudo, ya que permite ver los datos de manera muy clara, pero pueden llegar a verse de una altura bastante similar, per la profundidad permite un poco más de comparación respecto al de barras. La de pie permite ver el porcentaje que abarcan los datos de manera clara; sin embargo, dada la similar entre la cantidad de cada dato, se podría pensar que todos los años cuentan con el mismo porcentaje, apreciándose poca diferencia entre los mismos. Por último, la de línea ofrece una visión mucho más clara de las diferencias entre años, visualizando los incrementos entre uno y otro gracias a las conexiones entre puntos, siendo la mejor gráfica para ver comparativas entre los valores anuales.

3. Ganancias mensuales por condado:

Samuel: De acuerdo con los diagramas, se puede ver que, por una diferencia bastante significativa, el condado “POLK” es el que más ganancias ve reflejadas por venta de licor. Respecto al segundo con más ganancias, es decir “LINN”, la diferencia con “POLK” es de un poco más del doble, concluyendo que “POLK” es el condado donde más ventas de licor se hacen y muy probablemente donde más diste el precio de venta del costo original del licor, de forma que puede generar más utilidades. En la mayoría de otros condados, se ve que las ganancias no son tan destacables, sospechándose que no se logra ganar mucho de la venta de licor en bastantes condados. Añadido a esto, se puede evidenciar que la segmentación por meses es bastante uniforme, es decir que los valores se parecen bastante entre sí, denotando estabilidad a lo largo de los meses.

Lukas: Se puede ver que son muchos los condados que no tienen muchas ganancias por ventas de alcohol, lo cual puede deberse a que son condados más pequeños, y que por tanto no tienen una población tan grande interesada en la compra de alcohol ni muchas tiendas disponibles en las que comprarlo. Por ello, se concluye que es más rentable vender alcohol en condados grandes y concurridos, para que así se puedan realizar más cantidad de ventas, y de esta forma poder generar más ganancias. Además, podemos ver que no hay mucha variabilidad en las ganancias respecto a la segmentación por meses, por lo cual se puede tener seguridad en que

las ganancias no varían mucho según el mes, sino que más bien dependen de la parte de Iowa en la que se comercialice.

Daniel: Respecto a las gráficas, la de barras presenta un resumen completo de todos los datos y los valores que los mismos pueden tomar, mostrando la segmentación por meses de cada uno correctamente; sin embargo, pueden verse algo pegados entre sí, viéndose algo desordenados. En el de dispersión podemos ver mucho más claramente la variabilidad de los datos respecto al mes, y al no estar los puntos tan pegados se ve más ordenado, sin embargo, es más inexacto que el de barras en cuanto a los valores que toma cada punto. El de calor presenta una visualización interesante, ya que permite ver que tanta variabilidad presenta cada condado con los meses, esto según la intensidad de los colores, viendo que la mayoría tienen colores intensos, denotando estabilidad en meses. Por último, el diagrama polar presenta los datos de mayor a menor, siendo posible comparar de manera más efectiva la magnitud de los datos del análisis.

4. Ventas mensuales:

Samuel: De acuerdo con los diagramas, puede apreciarse que los meses en los que más hay ventas de licor es en mayo y en octubre, teniendo los dos casi el mismo número de ventas. Por otro lado, se aprecia que enero y febrero son los meses en los que hay menos ventas de licor. Hablando generalmente, si bien se puede ver una diferencia entre los meses según su número de ventas, la misma no es muy amplia, lo cual indica que los incrementos en ventas de alcohol, si bien se pueden apreciar, no son demasiado amplios, teniendo cierta consistencia en la cantidad de ventas entre meses.

Lukas: Si bien la diferencia de meses no es muy grande, tampoco es despreciable, por lo cual, si se desea maximizar los ingresos y ventas por alcohol estando en Iowa, es mucho más recomendable vender licor en mayo y octubre, ya que son estos meses en donde los datos apuntan a que se produzcan más ventas de alcohol. Así, se puede ver que la época del año en donde se venda alcohol influye en las ganancias obtenidas, ya que hay ciertos meses con más movimiento que otros.

Daniel: Respecto a las gráficas, la de barras y la de embudo permiten una buena visualización de los datos, con buen tamaño, anchura y diferencia visible entre los datos, siendo la de embudo un poco más efectiva en cuanto a las comparativas al presentar profundidad. Por otro lado, el diagrama de línea ofrece una mayor diferenciación entre las ventas de los meses, siendo posible observar las subidas o bajadas entre uno y otro. Por último, el de pie presenta los porcentajes

que abarca cada mes, pero es en el que menos diferencia puede evidenciarse entre los datos, viéndose todos muy similares entre sí.

5. Costo por categoría:

Samuel: De acuerdo con los diagramas, podemos ver que la categoría que más representa costos para las tiendas es "CANADIAN WHISKIES", seguida de "VODKA 80 PROOF", siendo las demás bastante similares en frecuencia, contando mayormente con valores considerables, pero no tan altos como las dos primeras. Se puede apreciar que este diagrama se parece bastante al diagrama de ventas por categoría, apreciándose que "CANADIAN WHISKIES" es mayor en costos, pero "VODKA 80 PROOF" es mayor en ventas, concluyéndose con estos dos casos que el costo y las ventas tienen una relación, ya que puede que se venda menos, pero puede ser mayor en costo.

Lukas: Comparando con el análisis de las ventas por categoría, se puede evidenciar que algunas que tienen menos valor en ventas, pero más valor en costos, por lo cual se puede sospechar que aquellos datos que tienen un valor mayor en costos que en ventas, se venden menos debido a que tienen un costo más elevado, explicando la relación entre ventas, costos y popularidad de un producto, viendo que, si cuesta menos, puede venderse más.

Daniel: Respecto a las gráficas, barras y embudo muestran los datos en su totalidad, viéndose un diagrama amplio y completo. Sin embargo, pueden verse algo pegadas las barras entre sí, lo cual genera cierta sensación de desorden y algunas dificultades menores en su lectura. En cuanto al de dispersión, se ve más ordenados que los anteriores, ya que los puntos tienen más distancia entre si de lo que tienen las barras, pero puede generar problemas en cuanto a la estimación de su valor, ya que los puntos son un poco más inexactos en la representación de valores que los anteriores.

6. Ingresos anuales por categoría de Vodka:

Samuel: De acuerdo con los diagramas, se puede apreciar que "VODKA 80 PROOF" es el tipo de vodka que genera más ingresos, seguido de "IMPORTED VODKA" y "VODKA FLAVORED"; sin embargo, estos están considerablemente alejados de los ingresos del primero. De igual manera, y viendo la segmentación por años, vemos nuevamente que los ingresos de cada categoría se mantienen estables por cada año, sin presentar mayores cambios. De acuerdo con datos de organizaciones de comercio en Iowa como lo es *"Iowa Alcoholic Beverages Division"*, el tipo de licor más consumido en Iowa es el vodka, lo cual se puede apreciar en estos datos, dados los altos valores que toman la mayoría de sus tipos.

Lukas: Viendo los datos, y teniendo en cuenta el supuesto anterior de que el vodka es el tipo de licor más consumido en Iowa, vemos que la mayoría de los tipos generan muchos ingresos. Esto sumado a que la categoría más consumida en general es "VODKA 80 PROOF", da la sospecha de que el vodka, efectivamente, es el tipo de licor más popular en Iowa, y es mucho más rentable vender vodka en Iowa. Por último, podemos ver que el año no influye mucho en cuanto a variabilidad de ingresos, viendo seguridad de que todo el tiempo, los grandes ingresos del vodka se mantienen estables.

Daniel: Respecto a las gráficas, la de barras muestra una visualización bastante completa de los datos, teniendo barras grandes y un buen orden en los mismos, viéndose muy clara la segmentación por años. La de dispersión 2D es bastante similar, presentando un mayor orden en los datos, pero siendo algo más imprecisa en cuanto a los datos exactos de cada una. Tambien se realizó uno de dispersión 3D, que permite ver con más claridad el estado de cada categoría respecto al año, siendo un diagrama muy interactivo y completo. El diagrama de calor permite ver con claridad cuales datos pueden variar más o menos según los años, viéndose que algunas varían más o menos según el año, esto de acuerdo a la intensidad del color que tengan. Finalmente, el diagrama polar permite ver de manera ordenada los datos de mayor a menor, ofreciendo otra forma de ordenar los datos y compararlos de mejor manera.

Conclusiones

Samuel: Este proyecto fue producto de un trabajo bastante extenso, esto debido a todo lo que implicó la consecución, visualización y análisis final de los datos.

Comenzando por la selección de la base de datos, que requirió de una búsqueda extensa, la cual tenía como fin conseguir una base de datos que tuviese varias columnas, llaves primarias, valores multivariados y algunos valores derivados, siendo estas características que cumple nuestra base de datos.

Una base de datos sobre venta de bebidas alcohólicas en una locación específica podría sonar en primera instancia como un proyecto simple y que no requiere de mucho análisis. Pero si se ahonda más en el mismo, se puede ver todo lo que implica la venta de una botella de alcohol, yendo desde los precios de compraventa, el lugar específico donde este se vende, el tipo de alcohol, el proveedor del alcohol y el nombre de la bebida, entre otros datos que hacen de este un proyecto complejo, y que tiene muchos datos posibles que analizar.

Añadido a esto, la complejidad en las condiciones que debe tener una base de datos nos permitió desarrollar una capacidad mayor de selección en cuanto a bases datos, aprendiendo a identificar cual es más o menos útil para crear un análisis mucho más completo y elaborado.

Lukas: En cuanto al diseño de la base, se tuvieron que hacer un numero grande de filtros, esto debido a que había datos que no estaban completos, que no eran del todo comprensibles o que simplemente no eran pertinentes para el estudio. Luego de esto se aplicó normalización a la tabla, sacando provecho de los diversos atributos de la misma, sacando una cantidad de entidades amplia, y que permitía un proyecto ordenado y bastante compuesto.

La subida de datos en PgAdmin4 se realizó mediante archivos csv, usando la función "COPY" para poder traspasar los datos a las tablas que con anterioridad se crearon en base al Excel normalizado, siendo posible subir más de 1 millón de registros en poco tiempo y sin mayores complicaciones.

Respecto a las conexiones con Dash mediante Python, esta fue la parte más compleja de hacer, ya que para ello se tuvo que manejar la librería "psycopg2", lo cual permitió tener una mayor comprensión y aprendizaje sobre conexiones entre plataformas, en este caso PgAdmin4 y Python, siendo un conocimiento bastante útil para proyectos a futuro.

Daniel: En cuanto al manejo de Dash, me llevo a adquirir bastantes conocimientos nuevos en el mundo de la programación, ya que en el mismo también se usa como agregado la librería "plotly.express", la cual se utiliza para gráficas, teniendo que aprender como graficar datos de una base de datos. De igual manera, también tuvimos que aprender lo básico de HTML, plataforma muy utilizada para diseño web, por lo cual estos conocimientos serán bastante aplicables para la carrera en un futuro.

Finalmente, la creación de escenarios de análisis y la discusión de las gráficas de los mismos nos permitió pensar en la situación problema desde otras perspectivas, aumentando la creatividad y los diferentes enfoques que se le pueden dar a un estudio o análisis. Respecto a las gráficas es bastante similar, ya que la realización de este proyecto nos permitió sacar conclusiones y deducciones sobre análisis que, vistos de otra manera, no habría sido posible analizar sin ayuda visual por la cantidad masiva de datos.

En general, fue un proyecto que implicó la adquisición de bastantes conocimientos, métodos, formas de pensar y analizar problemas, ya que, al ser nuestro primer acercamiento a la aplicación de bases de datos en contextos reales, permitió conocer todo el alcance que pueden llegar a tener las mismas si se hace su análisis y estudio de la manera correcta.

Este primer acercamiento sirve para darnos una idea de lo que podemos llegar a crear con bases de datos, dándonos las herramientas necesarias para en un futuro poder hacer proyectos más complejos y masivos, y así poder hacer uso de estos conocimientos en un mundo laboral y corporativo.

Anexos

Link del repositorio: https://github.com/Maicken052/Proyecto_Datos.git