Limpieza y transformación de datos

Contenido

Introducción

El análisis de datos es un proceso que implica recopilar, organizar e interpretar datos para responder preguntas o resolver problemas. Una de las herramientas que se pueden utilizar para el análisis de datos es Excel, un software de hoja de cálculo que permite a los usuarios manipular y visualizar datos de varias maneras. En este proyecto, demostraremos cómo usar Excel para realizar algunas tareas comunes de análisis de datos, como la limpieza de datos, la transformación de datos y la creación de paneles.

La limpieza de datos es el primer paso del análisis de datos e implica eliminar o corregir cualquier error, inconsistencia o valor faltante en el conjunto de datos. Esto puede mejorar la calidad y la precisión de los datos y facilitar el trabajo con ellos.

La transformación de datos es el siguiente paso del análisis de datos e implica cambiar la estructura o el formato de los datos para que sean más adecuados para el análisis. Esto puede implicar la creación de nuevas variables, la agregación o el resumen de datos, la división o combinación de datos, o la remodelación de los datos de formato ancho a formato largo o viceversa.

La creación de cuadros de mando es el paso final del análisis de datos, e implica presentar y comunicar los resultados del análisis de forma clara y concisa. Los paneles son informes interactivos que muestran métricas y tendencias clave mediante gráficos, tablas, segmentaciones, escalas de tiempo u otros elementos visuales y pueden ayudar a los usuarios a supervisar el rendimiento, identificar patrones o anomalías, comparar escenarios o tomar decisiones basadas en datos.

Eliminación de duplicados

El primer paso es eliminar los registros duplicados del conjunto de datos. Los registros duplicados son filas que tienen los mismos valores para todas o algunas de las variables. Pueden ocurrir debido a errores de entrada de datos, fusión de diferentes fuentes u otras razones. Dichos registros pueden afectar la calidad de los datos y dar lugar a cálculos e interpretaciones erróneos. Por lo tanto, debemos identificarlos y eliminarlos antes de continuar con el análisis. Esto garantizará que nuestros datos sean precisos y coherentes.

Columnas

Queremos hacer algunos cambios en el conjunto de datos para que sea más legible y útil. La primera columna, Id, permanecerá sin cambios porque es importante tener un identificador único para cada fila.

La segunda columna, Estado Civil, se modificará sustituyendo M y S por Casado y Soltero, respectivamente. Esto se puede hacer utilizando la función de búsqueda y reemplazo (ctrl + H). Esto facilitará al usuario la comprensión y el uso de esta columna.

La tercera columna, Género, también se modificará de manera similar, reemplazando M y F por Masculino y Femenino, respectivamente.

La cuarta columna, Ingresos, tendrá el formato de moneda en lugar de general, para mostrar la cantidad exacta de ingresos para cada fila.

La quinta columna, Edad, se agrupará en categorías basadas en rangos de edad. Utilizaremos una fórmula IF para asignar a cada fila una etiqueta de Adolescente, Mediana Edad o Senior, dependiendo de si la edad es menor de 31 años, entre 31 y 55, o mayor de 55. Esto nos ayudará a crear mejores visualizaciones sin tener demasiados valores de edad.

Creación de tablas dinámicas

Tabla 1:

Para crear un resumen de nuestros datos, utilizamos una tabla dinámica para mostrar los ingresos medios de diferentes grupos de clientes. Los agrupamos por género (hombre o mujer) y por si compraron una bicicleta o no (sí o no). De esta manera, pudimos ver cómo estos factores afectaban el nivel de ingresos de nuestros clientes.

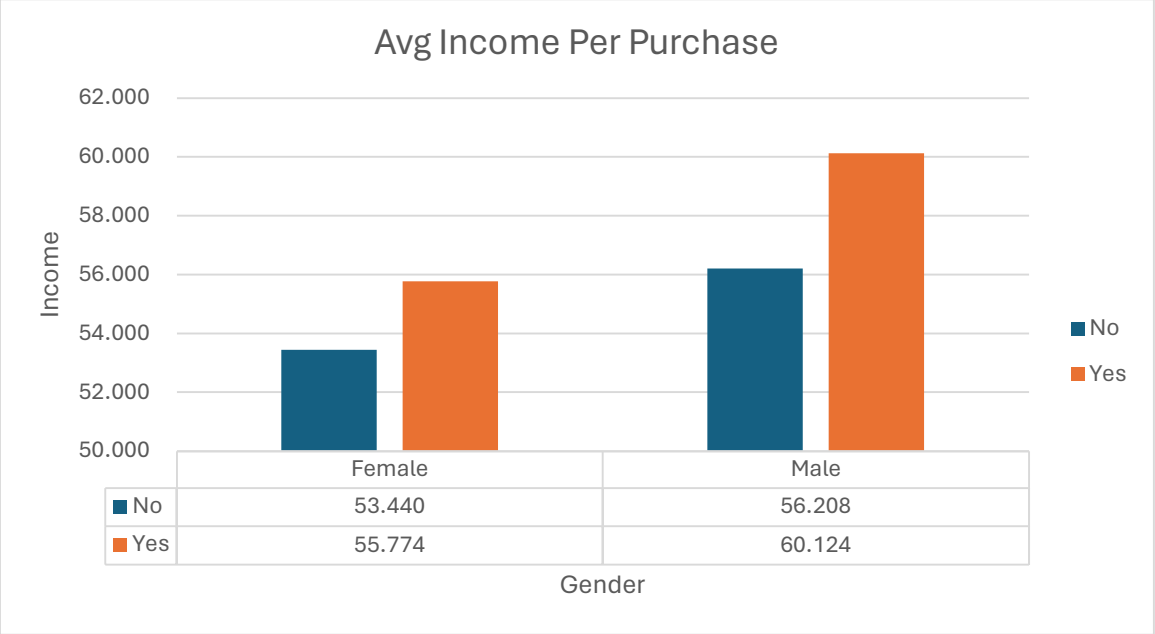
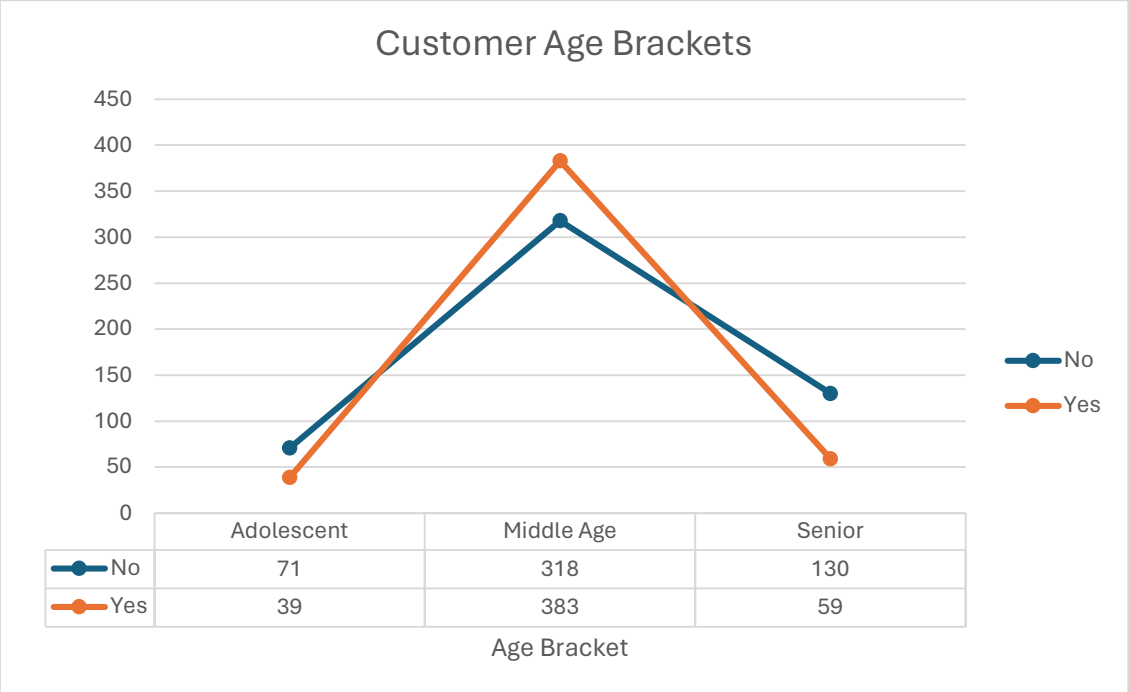


Tabla 2:

Uno de los factores que puede influir en la elección de una bicicleta es la distancia de desplazamiento. Las personas que necesitan viajar largas distancias pueden preferir bicicletas que sean cómodas, rápidas y eficientes en combustible. Por otro lado, las personas que utilizan bicicletas para viajes cortos pueden optar por bicicletas que sean fáciles de maniobrar, asequibles y ecológicas. Por lo tanto, es importante conocer la distancia media de desplazamiento de los compradores de bicicletas y cómo afecta a sus preferencias y satisfacción.

Tabla 3:

Para analizar la relación entre la edad y la compra de bicicletas, podemos utilizar una tabla dinámica que resume los datos por los grupos de edad que creamos anteriormente. La tabla dinámica mostrará el recuento de clientes que compraron una bicicleta y los que no, para cada grupo de edad. De esta manera, podemos ver cómo varía el comportamiento de compra de bicicletas en los diferentes rangos de edad.



Dashboard

El tablero se mejoró colocando las tablas dinámicas una al lado de la otra y eliminando las líneas de cuadrícula para una apariencia más limpia. Para permitir una mejor visualización y análisis, se insertaron segmentaciones basadas en la región, el estado civil y la educación. Los datos revelan algunas ideas interesantes.

Observaciones

Según nuestro análisis de datos, la mayoría de nuestros clientes pertenecen al grupo de mediana edad, mientras que los grupos de edad más jóvenes y mayores están menos representados. Esto sugiere que nuestros productos y servicios atraen más a las personas de entre 30 y 40 años que a las adolescentes o mayores de 60 años.

Las mujeres solteras compran más que los hombres solteros en los grupos de edad joven y mediana, pero esto cambia para el grupo de mayor edad, donde los hombres compran mucho más. Los hombres casados también compran más que las mujeres casadas para los grupos jóvenes y de mediana edad, pero no para el grupo de mayores, donde compran la misma cantidad. La principal conclusión de esto es que las mujeres solteras de mediana edad son las mejores clientas para nuestro negocio.

Los datos de ventas muestran que la región de América del Norte tiene las mayores compras de nuestros productos, seguida por los mercados de Europa y el Pacífico. Esto es sorprendente porque esperábamos que el mercado europeo fuera más receptivo a nuestras características ecológicas y sostenibles, ya que son conocidos por su conciencia y políticas ambientales. Sin embargo, parece que los consumidores norteamericanos valoran más nuestros productos, a pesar de sus estándares y regulaciones ambientales más bajos. Esto podría indicar una brecha entre la percepción y la realidad, o una diferencia en las preferencias y el comportamiento de los consumidores.

Según nuestros datos, los usuarios más frecuentes de nuestras bicicletas son aquellos que tienen una licenciatura. Este es un hallazgo sorprendente porque también ocupan el segundo lugar en términos de nivel de ingresos. Uno podría esperar que el grupo de ingresos más bajos dependiera de las bicicletas como un medio de transporte barato, o que el grupo de ingresos más altos disfrutara de las bicicletas como un lujo o un pasatiempo, pero nuestros números sugieren lo contrario.

Conclusiones

Según el análisis de datos, está claro que nuestros productos y servicios atraen más a las personas de entre 30 y 40 años. Las mujeres solteras de mediana edad son las mejores clientas para nuestro negocio. Por lo tanto, debemos centrarnos en crear campañas de marketing que se dirijan a este grupo demográfico. Las redes sociales son la mejor manera de llegar a las mujeres millennials, siendo Facebook e Instagram las plataformas más populares. Es importante tener en cuenta que debemos evitar estereotipar o generalizar nuestros mensajes de marketing. En su

lugar, debemos reducir nuestra demografía de marketing para ser lo más específicos posible y desarrollar contenido y mensajes que se dirijan a esta audiencia más refinada.

En términos geográficos, la región de América del Norte tiene las mayores compras de nuestros productos, seguida por los mercados de Europa y el Pacífico. Debemos seguir centrándonos en estas regiones y adaptar nuestras campañas de marketing a las necesidades y preferencias específicas de cada región. Por ejemplo, podríamos destacar las características ecológicas y sostenibles de nuestros productos en el mercado europeo, donde la conciencia y las políticas medioambientales son más frecuentes.

Por último, deberíamos plantearnos dirigirnos a aquellos que tienen una licenciatura, ya que son los usuarios más frecuentes de nuestras bicicletas. Este grupo ocupa el segundo lugar en términos de nivel de ingresos, lo que sugiere que valoran la conveniencia y los beneficios para la salud de andar en bicicleta por encima de otros modos de transporte. Podríamos crear campañas de marketing que destaquen los beneficios para la salud del ciclismo y cómo puede mejorar la calidad de vida.

En general, nuestras estrategias de ventas deben centrarse en la creación de campañas de marketing dirigidas que atraigan a grupos demográficos y regiones específicos, al tiempo que destacan las características y beneficios únicos de nuestros productos.