



**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
MONTERREY**

Análisis de datos para la toma de decisiones

Profesor: Cosme Ernesto Santiesteban Toca

Evidencia 1

Análisis estadístico de la base de datos

Equipo 20:

Diego Antonio Najera Rodríguez A01252951

Jesús Ernesto Gastelum Heras A01741236

Juan Marco Castro Trinidad A01742821

Mariana Peniche Castro A00835933

Fecha de entrega: 3 de septiembre de 2023

Parte 1

Parte 1

Base de datos seleccionada: [Da click aquí para ver base de datos](#)

- Por cada variable, identifique el tipo de la misma (discreta vs continua)

Variable	Tipo de variable
Entidad federativa	Cualitativa nominal
Periodo	Cualitativa politómico
Porcentaje de hechos presuntamente violatorios de derechos humanos registrados	Cuantitativa continua
Hechos presuntamente violatorios de derechos humanos registrados	Cuantitativa discreta
Total de hechos presuntamente violatorios de derechos humanos registrados	Cuantitativa discreta

- **Relevancia de las variables**

- Periodo: La variable del periodo es relevante, ya que nos dice los años que fueron tomados estos datos, y así nos podemos dar una idea del contexto del país en el 2016 y 2017. Esto es de gran importancia porque con esta variable se puede analizar cómo cambian los datos conforme pasa el tiempo, y nos servirá para el estudio al ir almacenando los datos de los siguientes años.
- Entidad Federativa: Las entidades federativas son importantes, puesto que nos dice el estado de donde vienen los datos, y esto puede explicar la razón por la cual los porcentajes son como son, como por ejemplo el contexto y la cultura de cada uno de los estados.
- Porcentaje: El porcentaje de hechos presuntamente violatorios de derechos humanos relativos a los derechos culturales es importante porque nos dice que tanta relevancia hay entre ambas violaciones de los derechos.

- Hechos violatorios humanos con culturales: Los hechos violatorios de derechos humanos relativos a los derechos culturales nos da una idea de la relevancia que hay entre la cantidad de las violaciones tanto como con los derechos humanos como con las culturales.
- Total de hechos violatorios humanos: El total de hechos violatorios de derechos humanos nos dice la cantidad de las violaciones cometidas en los periodos y estados dados y es importante para poder ver cuantos hechos violatorios humanos se han realizado en dicha entidad federativa y año.

- **Para variables discretas identifiqué la frecuencia de cada valor discreto**



- **¿Cuáles son los valores más frecuentes?**

Hechos presuntamente violatorios: los valores más frecuentes se encuentran entre el 0 al 99.

Total de hechos presuntamente violatorios: los valores más frecuentes se encuentran entre el 0 y el 4999.

- **¿Qué conclusiones puede derivar de lo anterior?**

Con base en los datos analizados anteriormente, pudimos concluir que en ambos años (2016 y 2017) la mayoría de los hechos presuntamente violatorios de derechos humanos registrados a nivel nacional por año en cada estado de México se encuentran entre los 0 y 99 casos, mientras que el total de hechos registrados es alrededor de 0-5000 en cada estado en total. Esto nos quiere decir que, entre el 2016 y el 2017 los casos que se han dado forman tan solo un pequeño porcentaje del total y no se han dado en gran cantidad; esto lo podemos ver también en la columna de porcentajes que nos muestra a mayor detalle cada estado.

- Para variables continuas, seleccione un número de intervalos y agrupe los valores observados.



- Determine la frecuencia observada sobre los intervalos antes definidos
 - ¿Cuáles son los intervalos más frecuentes?

A partir de lo observado, los intervalos más frecuentes están entre el 0 y el 0.99%.

- ¿Qué concluye de lo anterior?

Podemos concluir que esto es algo bueno debido a que en los años 2016 y 2017 en México, no se dieron muchos casos de violación a los derechos humanos, ya que se muestran unos porcentajes muy bajos. Podemos ver que hay excepciones, como lo son los estados de Veracruz, Tlaxcala y Tamaulipas que llegan a porcentajes de entre el 7 y el 8% en tan solo un año lo cual es preocupante.

- Preguntas:

- ¿Por qué esa pregunta y variables asociadas son relevantes?

- Es importante tener en cuenta estas preguntas, ya que por medio de los datos que nos entregan es posible conocer cuál es el estado que más está causando problemas a los ciudadanos, pero no solamente tiene que ser por estado, sino que también puede llegar a existir una prevalencia por sector de México, que es lo que responde la cuarta pregunta por ejemplo. Tener en cuenta el año en que sucedió nos puede ayudar a explicar el porqué sucedieron estas violaciones a los derechos y verificar si hubo algo de por medio que las causara en cierto o ciertos

estados, esto con el propósito de mejorar la atención y servicio que se le ofrecen a la población sin necesidad de perjudicarlos.

■ **¿Cuál puede ser el impacto de comprender las variables, dado el contexto del que se extrajo la base de datos?**

La importancia de conocer las variables y ponerlas en el contexto es de suma importancia y relevancia para su estudio, ya que eso nos permite comprender mucho mejor lo que estamos analizando y nos permite poder hacer un estudio mucho más profundo de nuestro tema. Al no comprender las variables, no podremos formularnos buenas preguntas y no nos permitirá poder presentarlas, puesto que ni siquiera se nos ocurrirá qué poder hacer con ellas. Es necesario analizar primero el contexto de lo que estamos estudiando, después analizar las variables y posteriormente ya empezar a analizar los diferentes datos que tenemos para poder convertir esos datos en información.

● **Por cada variable, determine: Media aritmética, Mediana, Moda y Desviación estándar**

- Porcentaje de hechos presuntamente violatorios de derechos humanos
 - Media aritmética: 1.56
 - Mediana: 0.7
 - Moda: 0
 - Desviación Estándar: 2.16
- Hechos
 - Media aritmética: 75.65
 - Mediana: 13
 - Moda: 0
 - Desviación Estándar: 180.47
- Total
 - Media aritmética: 4677.29
 - Mediana: 1824.5
 - Moda: 586
 - Desviación Estándar: 6615.89

- **¿Qué concluye de los valores calculados en los dos puntos previos?**
 - Analizando los diferentes datos, que sacamos en los puntos previos, podemos darnos cuenta de que, en nuestras variables, la desviación estándar es bastante alta, lo que nos quiere decir que tenemos muchos datos muy dispersos, es decir, que tenemos datos muy bajos y otros datos muy altos. Esto lo podemos ver representados por ejemplo en el total de hechos violatorios, en donde vemos que la mediana y la media aritmética son muy diferentes, por lo que esto nos dice que tenemos muchos datos muy bajos y algunos datos muy altos, lo que hace que nuestra media aritmética suba. Podemos confirmar dicha diferencia también analizando nuestra desviación estándar, el cual podemos ver que, por el hecho de tener algunos datos tan altos, afecta nuestra desviación estándar haciéndola un valor más alto. En conclusión, podemos decir que, el hecho de analizar datos como la media, mediana, moda y desviación estándar, que se pueden ver datos simples, nos ayuda a poder comprender el centro y la dispersión de nuestros datos, para poder comprenderlos un poco mejor.

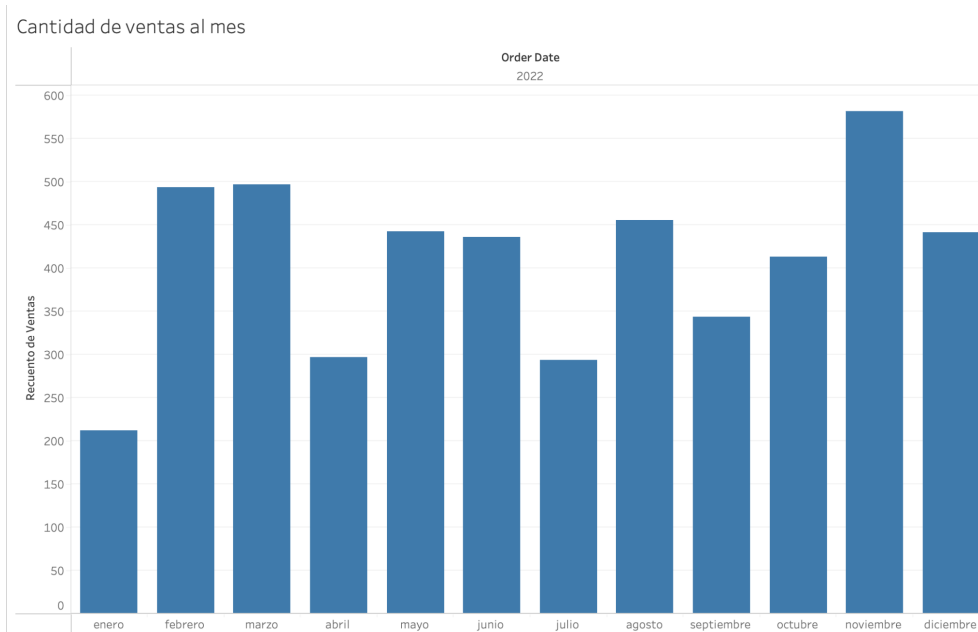
Parte 2

Parte 2

Instrucciones: Para una variable discreta, derive una representación visual que muestre frecuencia de los valores, valores de centralidad, distribución.

Variable discreta elegida: Fecha

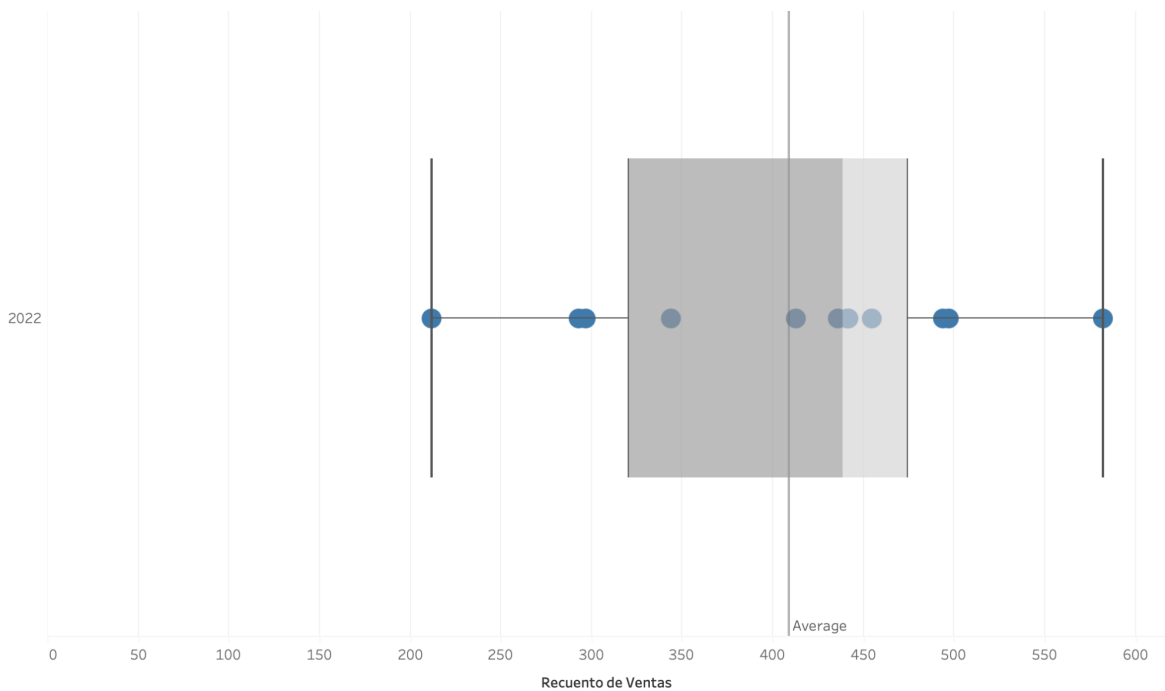
Frecuencia de los valores:



Explicación: En este caso, nosotros elegimos de variable discreta la fecha y la dividimos por los diferentes meses del año. Ahí mismo, nosotros podemos observar que mediante el uso de un histograma pudimos representar diferentes medidas de la variable fecha o “order date” donde concluimos que el mes de noviembre es el que tuvo más ventas de la Tec Store durante el año de 2022, seguido por marzo y febrero. También se puede observar que es una distribución multimodal debido a que no hay sigue un patrón en particular, por lo que las ventas pueden variar en gran cantidad durante el año.

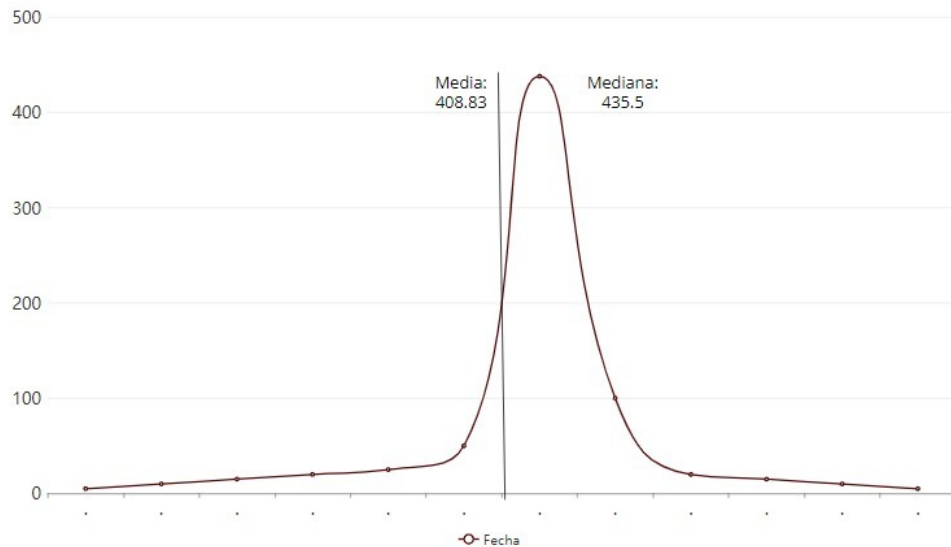
Valores de centralidad: (media/promedio, desviación estándar, mediana, moda)

- Media: 408.83
- Mediana: 438.5 (entre junio y mayo)
- Moda: noviembre del 2022 (11/2022)
- Desviación estándar: 99.167



Explicación: El diagrama de cajas y bigotes nos indica que la cola de la distribución se encuentra más hacia la derecha debido a que el cuartil 2 está más cerca del cuartil 3. Finalmente, el valor de la desviación estándar nos indica que los valores pueden variar de la media por 99.167.

Distribución de los valores:



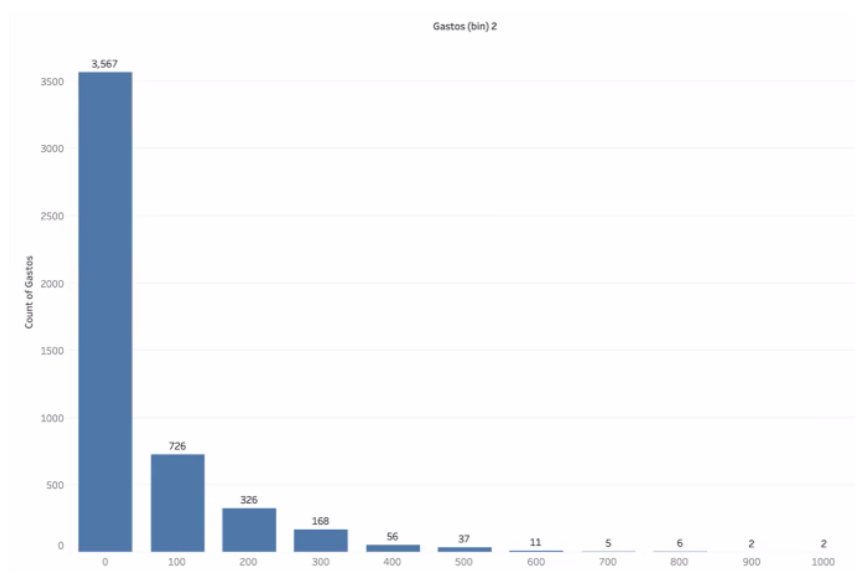
Explicación: Analizando la distribución de los valores, podemos darnos cuenta de que la mayoría de las ventas, en este caso, fueron casi proporcionales la primera mitad del año y la segunda mitad del año. Sin embargo, como podemos observar, podemos ver que tenemos ligeramente hacia la derecha nuestra mediana, lo que significa que nuestros datos estuvieron más inclinados hacia la segunda mitad del año, lo que significa que tuvimos ventas más altas en los meses del segundo semestre del año. Esto se puede ver reflejado por la temporada del año, por momento/eventos del Tec o por muchos otros diferentes factores.

Segunda parte:

Instrucciones: Para una variable continua, determine un conjunto de intervalos y agrupe los valores observados, derive una representación visual con frecuencia, valores de centralidad y distribución.

Variable continua elegida: Gastos

Frecuencia de los valores agrupados en conjuntos de intervalos:



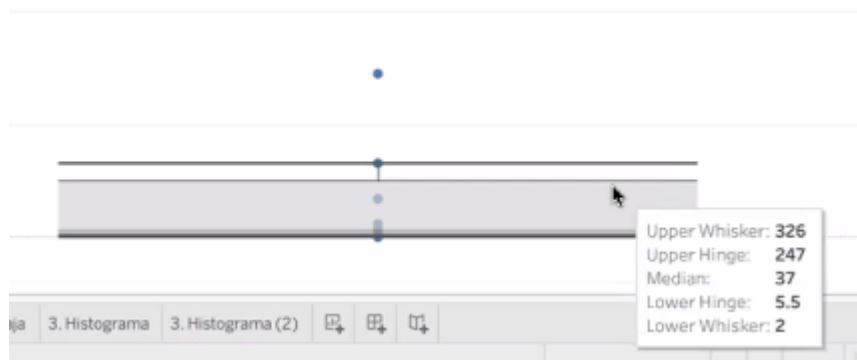
Explicación: Nosotros elegimos la variable de gastos y la agrupamos en conjuntos de intervalos de 100 en 100. Después de agrupar y tener nuestros conjuntos, nosotros realizamos este histograma de frecuencias para poder visualizar cuantos datos entraban en cada intervalo y como esto se veía reflejado. Al momento de analizarlo, podemos observar que la mayoría de los datos están fuertemente ubicados al inicio, pero podemos observar como otros datos mucho mayores, en cuestión de gastos, también existen, pero que hacen que lleguen a poder afectar nuestros valores de media y mediana.

Valores de centralidad: (media/promedio, desviación estándar, mediana, moda)

Sacados en Excel:

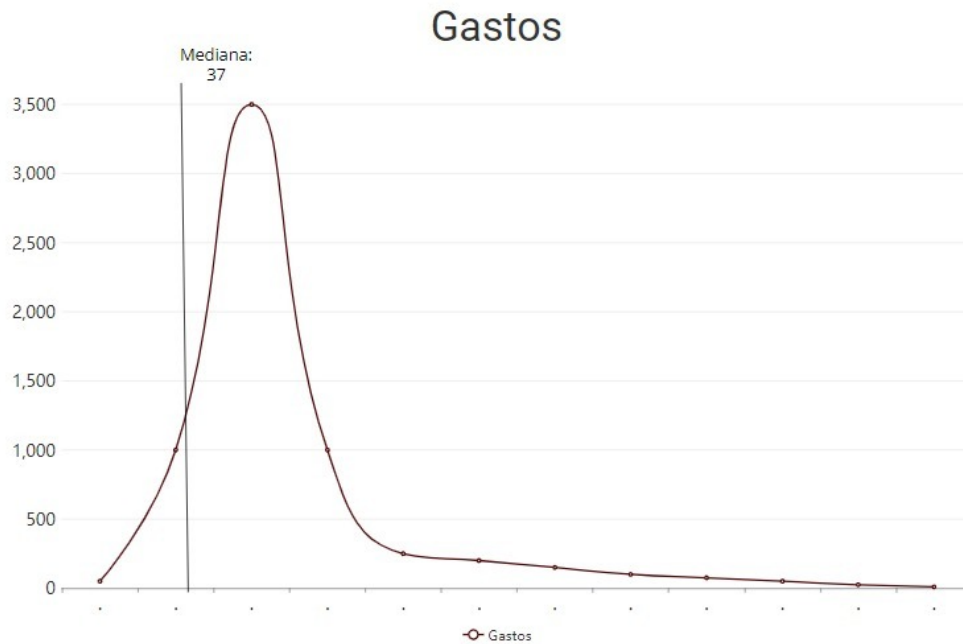
- Mediana = 90
- Promedio = 97.69395128
- Desviación Estándar = 109.69
- Moda = 90

Sacados en Tableau:



Explicación: Aquí, es interesante lo que pasa. Nosotros, si realizamos los cálculos de la moda, mediana, media en Excel, nos dan los valores anotados previamente. Sin embargo, al sacarlo en Tableau, vemos que nuestros datos cambian. Esto es porque en Tableau podemos darnos cuenta de que hay varios valores atípicos que desvirtúan nuestros resultados, por lo que Tableau no los contempla. Al analizar el diagrama de caja podemos observar que la mayoría de nuestros datos, como lo habíamos observado en nuestra gráfica de frecuencia, estan altamente inclinados hacia el inicio. Por último, analizar y entender nuestros cuartiles, sin considerar los valores atípicos, nos ayuda a tener números muchos más sólidos y observar que realmente está pasando con nuestros gastos.

Distribución de los valores:



Explicación: Analizando por último nuestra gráfica de dispersión, podemos reafirmar por última vez de que nuestros datos están alejados completamente de nuestro centro, lo que nos dice que nuestros gastos son mayormente bajos, pero que también tenemos ciertos valores muy altos que son realmente muy mínimos. Es por eso, que el analizar también, junto con los otros diagramas, nos ayuda a comprender entre qué rangos están nuestros gastos, para poder entender mucho mejor que es lo que está pasando con Tec Store.

Tercera Parte:

Instrucciones: Determine los cuartiles de una variable continua. Represente los cuartiles en un histograma y en un diagrama de caja.

Variable continua elegida: Ventas

Histograma:

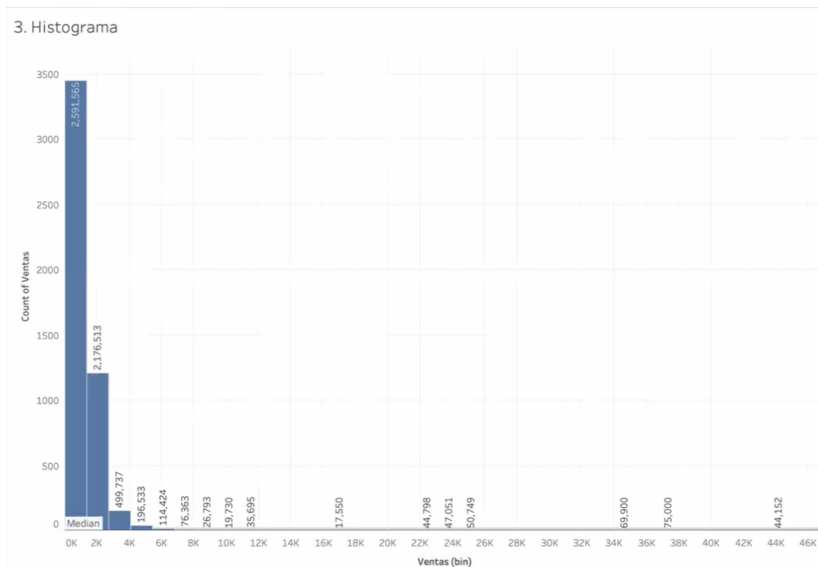
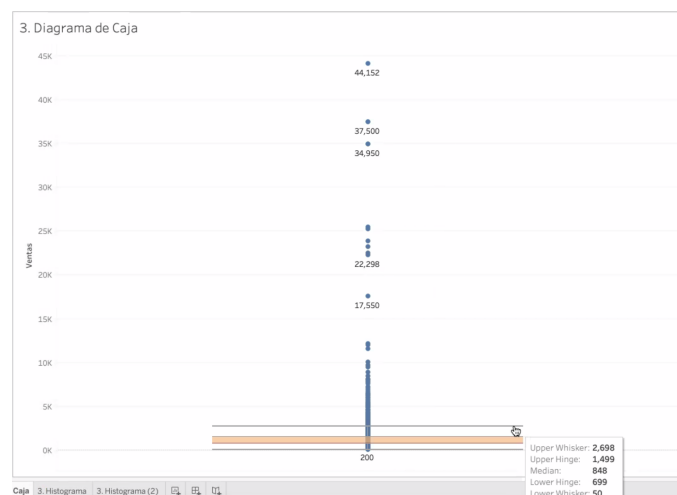


Diagrama de caja:

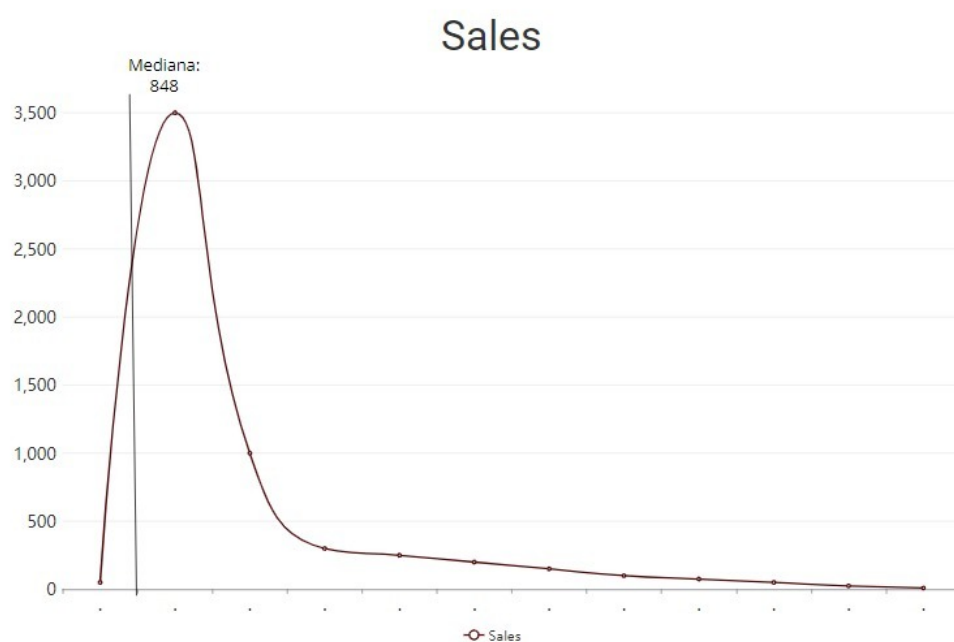


Cuartiles:

- Extremo Inferior: 50
- Cuartil Inferior: 699
- Mediana: 848
- Cuartil Superior: 1,499
- Extremo Superior: 2,698

Explicación Ambas: Analizando nuestro histograma y diagrama de caja, podemos observar que paso algo similar a nuestros datos. Primeramente, observamos fácilmente de que nuestros datos están concentrados en nuestra primera mitad, por lo que ahí es donde entran la mayoría de nuestras ventas. Sin embargo, el hecho de tener otros valores muchos más grandes, nos habla de si podemos llegar a tener ventas muchos más grandes, pero al analizarla podemos ver que claramente no significa que en realidad vendamos eso. Agregando a lo anterior, también podemos observar que en nuestro diagrama de caja hay puntos afuera de nuestros mínimos y máximo valor. Estos valores representan esos valores atípicos que lleguen alterar nuestros datos fuertemente, pero es por eso que al hacer el diagrama de caja, podemos observar entre qué mínimo y qué máximo, y en donde está más concentrado nuestros datos

Extra: Distribución de los valores



Pregunta Final:

¿Qué ventajas/desventajas puede identificar de cada una de estas representaciones?

- **Histograma**
 - Las ventajas de utilizar el histograma son las siguientes:
 - Nos permitió representar de forma visual la distribución de las ventas anuales que tuvo la Tec Store.
 - Ayuda a realizar una comparación y análisis eficaz y rápido para identificar los meses con más ventas y buscar la causa de porque en ciertos periodos se vende más que en otros.
 - Lo podemos utilizar para presentarlo como un gráfico definitivo en algún reporte.
 - Mientras que las desventajas de este mismo son:
 - Puede no ser muy eficaz si queremos utilizarlo para el detallado más específico sobre los valores de ventas individuales de cada mes.
 - No puede ayudarnos como herramienta de predicción por la frecuencia de ventas que sería útil al manejar una página en línea.
- **Diagrama de caja**
 - Por otro lado, aquí se explican las ventajas del diagrama de caja:
 - Una de las ventajas es que en este diagrama se muestra la mediana, cuartiles y posibles valores atípicos que ayuda con la comprensión de la distribución de ventas mensuales.
 - Ayuda a identificar las tendencias de ventas.
 - Es más sencillo realizar una comparación mensual.
 - Y las desventajas de este diagrama son:
 - Los datos son muy poco detallados, ya que no se puede observar en esta gráfica la distribución de los datos así como en el histograma.
 - Puede llevar a una menor comprensión por la simplificación excesiva que se muestran en los datos de esta gráfica.

En conclusión las dos gráficas son útiles para realizar este tipo de estudios, pero cada una cuenta con datos o métricas específicas que ayudan a la retroalimentación de los datos obtenidos a lo largo del año y las posibles estrategias para mejorar estos números y el desempeño de en este caso la página web del Tecnológico de Monterrey “TecStore”.

Parte 3

Parte 3

Al momento de estar trabajando con diferentes bases de datos, el hecho de hacer diferentes gráficas y el hecho de hacer diferentes dashboards, nos dio una perspectiva muy grande y un conocimiento vasto de cómo el análisis estadístico nos ayuda a nosotros a poder estudiar, decidir y pronosticar diferentes cosas. A lo largo de las actividades, pudimos darnos cuenta de cómo el hacer el análisis estadístico nos ayuda a poder analizar lo que está pasando en nuestra empresa, poder analizar los datos para posteriormente convertirlos en insights e información que nos ayude a poder mejorar lo que sea que estemos estudiando o simplemente entender más a fondo el porqué pasan las cosas que están pasando en nuestra empresa. Un ejemplo de lo anterior es en la parte de logística.

Tomando como referencia la base de datos de Tec Store, al momento de analizar nuestras diferentes variables y al hacer las gráficas, pudimos observar que productos son los que más se vendían, pudimos ver quienes eran nuestros mayores clientes, donde nos compraban más, a que hora y que día o temporada lo hacía, entre muchas otras cosas. De esa manera, nosotros podemos llegar a tomar diferentes decisiones en cuestiones de logística y requerimientos de recursos mucho más inteligentes que buscan que la empresa suba más sus ventas, clientes, órdenes y ganancia. Al analizar todos estos elementos nosotros podemos llegar a decidir si es necesario producir más de un producto, prescindir de aquellos no tan populares, buscar el porqué algún producto tiene ventas más altas y buscar seguirlo explotando o incluso analizar porque un producto tiene ventas bajas y decidir que se puede hacer para mejorar, entre muchas otras decisiones que van saliendo gracias al tener un análisis estadístico. Y lo más interesante de todo es que el análisis estadístico es aplicable para cualquier base de datos o cualquier contexto en el que estemos trabajando y busquemos analizar y/o mejorar algunas cosas.

En resumen, el análisis estadístico resulta estar más relacionado de lo que parece con la toma de decisiones en la vida diaria, ya que esta nos da la oportunidad de evaluar las cosas desde una perspectiva fundamentada con una base objetiva y cuantitativa. Este tipo de análisis lo utilizamos todos los días, como por ejemplo en la evaluación de los riesgos que pueden surgir al tomar una decisión, o también al analizar la forma más objetiva, rápida y eficiente de realizar un trabajo. Por lo que en conclusión, el análisis estadístico es una de las herramientas que nos facilitan la vida al ayudarnos a realizar decisiones más acertadas y

estratégicas en una amplia variedad de áreas y aplicaciones, tanto como lo es en los negocios, como en cualquier área.