Unidad 3 - Análisis exploratorio de datos: Medidas de Resumen

Fundamentos de ciencia de datos

Ejercicio N°1

El dataset “alimentos.csv” fue elaborado por una clínica de nutrición que suministró a los/las pacientes una lista de alimentos permitidos con sus respectivos contenidos calóricos, también se detalló el tipo de alimento del que se trataba (fruta, verdura, fruto seco o elaborado) y el tipo de vitamina que aportaba cada alimento (A, B o C).

Además, la nutricionista a cargo del estudio lleva una planilla de control de evolución de 50 pacientes (”pacientes.csv”) en el que registra la edad, el sexo, la altura, el peso inicial y el peso final de cada paciente luego de seguir un plan de dieta por una cierta cantidad de tiempo, el cual también fue registrado en el campo “tiempo de tratamiento”.

Importe los dataset al entorno de trabajo.

1. Realice una descripción general del conjunto de datos que incluya la descripción de la información brindada por cada columna, el tipo de datos de cada una , el numero de registros y el rango de las mismas.
2. Calcule la media, la moda, la mediana, la varianza, la desviación estándar y la MAD del campo “aporte_calorico_kcal”.

3. Para el mismo campo, calcule Q1, Q3 y el rango intercuartil. Luego calcule los percentiles 25, 50 y 75, ¿coinciden con algun valor?.
4. Visualice la distribución de los valores en un histograma, marque en el mismo con una línea roja la moda, la media y la mediana.
5. Visualice la distribución de los datos en un boxplot, identifique y marque en el gráfico: la mediana, Q1, Q3 y el rango intercuartil.
6. Detecte valores atípicos en el campo “aporte_calorico_kcal” y elimínelos, vuelva a realizar un boxplot, ¿Qué cambios observa con respecto al grafico del punto e?.
7. Calcule la media, la mediana y la moda del campo “aporte_calorico_kcal” pero esta vez agrupandolo por “tipo_de_alimento”. Realice un boxplot en el que se observe la distribución de los aportes calóricos para cada categoría de “tipo_de_alimento”. ¿Cuál es la categoría de alimentos que parece aportar menos calorías? ¿Qué categoría de alimentos aporta valores calóricos más variables? ¿Cuál es menos variable? ¿Qué medida observa para determinarlo?.
8. Utilizando los datos de los/las pacientes agregue una columna a la tabla en la que se calcule la variación del peso corporal para cada paciente (calculado como [peso_final_kg]-[peso_inicial_kg]). Mediante gráficos boxplot muestre las diferencias en la variación del peso corporal para cada género. ¿En que género funciono mejor el tratamiento? ¿Existen valores atípicos en la distribución de alguno de los géneros?.
9. Realice un boxplot que muestre las diferencias en el tiempo de tratamiento entre los géneros, calcule el promedio, la mediana, la moda y la desviación estandar para ese campo.

Ejercicio N°2

Explore el conjunto de datos “winequality-red.csv”:

1. Realice una descripción general del conjunto de datos que incluya la descripción de la información brindada por cada columna, el tipo de datos de cada una , el número de registros y el rango de las mismas.
2. Realice una limpieza de datos que incluya el manejo de valores faltantes, duplicados y/o erróneos.
3. Explore la distribución de los valores en las columnas numéricas del dataset a través de un boxplot. Construya una tabla con el nombre de la variable en la

primera columna, la media en la segunda, la moda en la tercera y la mediana en la cuarta. Agregue una quinta columna con la desviasión estandar.

4. Realice un histograma para observar la distribución de datos en los campos acidez volátil y pH, calcule el coeficiente de variación para cada uno. Luego, compare los resultados obtenidos para determinar cuál de estas características presenta mayor variabilidad relativa.

Ejercicio N°3

Importe y explore el conjunto de datos “titanic.csv”:

1. Realice una descripción general del conjunto de datos que incluya la descripción de la información brindada por cada columna, el tipo de datos de cada una , el número de registros y el rango de las mismas.
2. Realice una limpieza de datos que incluya el manejo de valores faltantes, duplicados y/o erróneos.
3. Calcule la media, la mediana y la desviación estándar de la edad de los/las pasajeros/as que murieron y sobrevivieron para cada clase. Realice un boxplot que muestre la distirbución de edades para cada grupo(murieron/sobrevivieron) dentro de cada clase .¿En qué clase las edades de las personas que sobrevivieron fueron mas variables?.¿Cuál fue la edad de la persona más joven que sobrevivió en tercera clase?.
4. Calcule el promedio, la moda, la mediana, la desviación estandar y el rango intercuartil del precio del pasaje para cada clase. Muestre mediante histogramas y boxplots la variacion de los precios del pasaje por clase.
5. ¿Que medida de resumen calcularia si quisera conocer cuales fueron los/las pasajeros/as que pagaron un 25% mas caro el precio del pasaje que el resto ? ¿Estos pasajeros/as viajaban con otras personas (padres/madres, hijos/hijas, hermano/as y/o conyugues). Construya una tabla con los nombres de estos/as pasajeros/as en una columna, el numero total de personas vinculadas a ellos/as que se encontraban en el barco en la otra y la cuidad en la que embarcaron en la tercera. ¿Con cuantas personas en promedio viajaban?¿En que puerto embarcaron la mayoria?.

Ejercicio N°4

Hasta ahora trabajamos con las variables numericas interpretandolas como continuas, pero ¿que pasa si queremos resumir la información de las columnas que

contienen datos categóricos o cuantitativos o quisieramos considerar variables numéricas pero como discretas?.

1. Utilizando el dataset titanic.csv, calcule la cantidad de pasajeros/as que viajaron en cada clase, asi como el porcentaje, la frecuencia relativa, la frecuencia relativa acumulada, la frecuencia absoluta y la frecuencia absoluta acumulada para de cada grupo. Presente los resultado en una tabla. Ilustre las diferencias en un histograma. ¿Qué puedes decir acerca de la distribución de pasajeros/as en las diferentes clases?.
2. Calcule la cantidad y el porcentaje de pasajeros que subieron al barco acompañados por 1 o mas personas de su grupo familiar. Construya un histograma para observar la distribucion de los datos. ¿Esta considerando a la variable “cantidad de acompañantes” como una variable continua o discreta?.
3. Construye una tabla de contingencia cruzando las variables "survived" y "class". ¿Qué proporción de personas de cada clase sobrevivieron y murieron al naufragio del Titanic?. Mustre los resultados en un gráfico de barra.
4. Elabora un histograma que muestre la cantidad de personas de genero masculino y femenino que sobrevivieron y murieron por intervalo de edad, los intervalos de edad deben ser: 0-18, 19-35, 36-56 y >57.
5. Construye una tabla de contingencia que muestre la cantidad de personas de genero femenino y masculino que murieron y sobrevivieron por clase económica. Muestre los resultados en un gráfico de barra.

Ejercicio N°5

Vuelva a cargar el dataset “alimentos.csv” y construya una tabla de frecuencias que resuma la cantidad de alimentos que aportan vitamiena A, B y C. Calcule la frecuencia relativa y absoluta en cada grupo. ¿Que proporción de alimentos aportan vitamina A?¿ que porcentaje de alimentos aportan vitamina C?. Presente los datos en un gráfico de barras en el que en el eje “y” se muestre el porcentaje de alimentos y en el eje “x” el tipo de vitamina que aporta.

Ejercicio N°6

Utilizando el set de datos winequality-red.csv

1. Construye la matriz de covarianza entre todas las variables numéricas y grafícala.

2. Encuentra los 5 pares de variables que tienen la mayor covarianza positiva y negativa en el set de datos, grafíquelas y describa en palabras como se relacionan.
3. Construye la matriz de correlación de Pearson entre todas las variables numéricas y grafícala.
4. Encuentra los 5 pares de variables que tienen la mayor correlación positiva en el set de datos, grafique la relación y describa en palabras como se relacionan. Repita la operación, pero con los 5 pares de variables que tienen la mayor correlación negativa. ¿Las variables con la mayor covarianza (positiva o negativa) coinciden con las de mayor correlación? ¿Pueden existir dos variables que tengan un alto índice de covarianza pero no estén correlacionadas entre sí?
5. Calcule la matriz de correlación de Spearman y compárela con la matriz de Pearson construida en el punto d). ¿Qué puede concluir a cerca de la forma en que se correlacionan las variables? ¿Qué información podría aportar al análisis construir la matriz de correlación de Spearman?

Ejercicio N°7

Utilizando el dataset “calidad_producto.csv”

1. Genere la matriz de correlación y calcule el coeficiente de correlación de Pearson entre las dos variables.
2. Calcule el coeficiente de correlación de Spearman, ¿son similares los resultados?
3. Realice un gráfico de dispersión de un variable vs. la otra, observe los puntos. Busque y remueva valores atípicos en el dataset y vuelva a calcular ambos coeficientes. ¿que observa?

Ejercicio N°8

Utilice el dataset “estación_meteorologica.csv”

1. Calcule y grafique las matrices de correlación y covarianza.
2. Identifique las variables más correlacionadas y grafíquelas una vs. la otra en un gráfico de dispersión. Luego grafique cada una a lo largo del tiempo (para todas las fechas).
3. Identifique las variables menos correlacionadas y grafíquelas una vs. la otra en un gráfico de dispersión. Luego grafique cada una a lo largo del tiempo

(para todas las fechas).

¿Qué puede concluir de los gráficos del punto 2 y 3?