Práctica Individual
Semestre 2025-EJ

1 Instrucciones

- La práctica es individual y debe entregarse el **viernes 11/04/2025** antes de las **23:59 h**.
- Debe escribir un reporte y hacer algo de código en Python.
- El reporte (un archivo PDF) debe describir el análisis de los datos y sus estadísticas. **Describa con sus propias palabras qué perspectivas y/o conocimientos puede extraer del análisis de las estadísticas**.
- El archivo PDF llevará como nombre **practica.pdf**.
- Escriba un código en Python para cada punto, y guárdelo con el nombre correspondiente: **programa_01.py, programa_02.py, programa_03.py**.
- Escriba su nombre completo, el nombre del curso y la fecha en la primera página del reporte, y como encabezado en cada archivo .py.
- Comprima el archivo PDF y los archivos .py en un único archivo **.zip** con el nombre **practica.zip**.
- Comprima los archivos directamente en el .zip. **No ponerlos en un directorio**.
- Previo al envío, compruebe **el tipo, el formato y los nombres de los archivos**. En caso de encontrar algo incorrecto, se restarán puntos de la calificación final.

2 Descripción

Adjuntado a este documento, se encuentra un archivo CSV llamado **survey_results.csv** que contiene los datos de una encuesta realizada a usuarios de Stack Overflow en 2024 que eran codificadores profesionales. Estos datos ya fueron previamente filtrados, conservando las respuestas de **21,784** usuarios. Los datos siguen una forma estructurada definida por las siguientes 6 variables (el nombre exacto de las variables en el archivo es el que está entre paréntesis):

- Nivel educativo (**EdLevel**).
- Años de experiencia de codificación profesional (**YearsCodePro**).
- Ocupación (**DevType**).
- País (**Country**).
- Lenguaje(s) de programación con experiencia (**LanguageHaveWorkedWith**).
- Salario anual en USD (**ConvertedCompYearly**).

2.1 Descripción de las variables

2.1.1 EdLevel

Para esta variable categórica, los grados académicos están en el siguiente orden creciente:

1. Primary/elementary school.
2. Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.).
3. Some college/university study without earning a degree.
4. Associate degree (A.A., A.S., etc.).
5. Bachelor's degree (B.A., B.S., B.Eng., etc.).
6. Master's degree (M.A., M.S., M.Eng., MBA, etc.).
7. Professional degree (JD, MD, Ph.D, Ed.D, etc.).

2.1.2 YearsCodePro

Es una variable numérica con un rango entre 1 y 50 años.

2.1.3 DevType

En este caso, esta variable tiene un total de 33 categorías, correspondientes a los trabajos actuales de los codificadores profesionales. La siguiente lista muestra algunos valores que podrá encontrar al realizar el análisis:

- Academic researcher.
- Blockchain.
- Developer, AI.
- Developer, back-end.
- Engineering manager.

- Hardware Engineer.
- System administrator.
- Security professional.

2.1.4 Country

Por otro lado, la variable **Country** cuenta con 165 categorías, que corresponden a diferentes países del mundo.

2.1.5 LanguageHaveWorkedWith

Para esta variable, los usuarios podían escoger más de una opción. Por ejemplo, C; C++; JavaScript; Python. En este caso, para las estadísticas, el mismo usuario contará para cada lenguage que haya escogido.

2.1.6 ConvertedCompYearly

Es una variable numérica con un rango que abarca desde 1 hasta millones.

3 Análisis a realizar

La práctica consiste en realizar 3 ejercicios con algunos subpuntos descritos a continuación sobre el procesamiento y análisis de los datos contenidos en el archivo CSV. Para cada ejercicio, debe escribir un código en Python. Además, debe describir en el reporte los resultados obtenidos con el código.

En cada ejercicio, encontrará subpuntos que corresponden a codificar y otros a escribir algo en el reporte. En el código y el reporte, escriba la frase correspondiente al punto que está codificando, analizando o explicando. Finalmente, reporte **las estadísticas, métricas, gráficos y distribuciones necesarias** para explicar su respuesta y el porqué de la misma, y describa con sus palabras cuáles son sus **hallazgos, percepciones y conclusiones** sobre el subpunto.

1. **Análisis por país.**
   (a) **Código:** Encuentre los 10 países con más respuestas y haga un gráfico de barras de las frecuencias de respuestas de esos países, es decir, cuántas personas hay por país. Si México no está entre los 10 primeros, incluya su frecuencia en el gráfico.
   (b) **Código:**
      i. Seleccione los 3 países con más respuestas e incluya a México si no está ahí.
      ii. Calcule el resumen de los cinco números (mínimo, primer cuartil, mediana, tercer cuartil y máximo), la media, y la desviación estándar para el salario anual para cada uno de los países seleccionados. Haga los diagramas de caja e histogramas (con 10 bins) correspondientes para cada país.

(c) **Reporte:**

i. Copie las estadísticas y gráficos calculados anteriormente por país (sólo los seleccionados). Organice las estadísticas y los gráficos para poder comparar fácilmente los datos de los distintos países (por ejemplo, colocando los gráficos uno al lado del otro). De acuerdo a las estadísticas calculadas y el número de personas por país, responda lo siguiente:

A. ¿Cuál de los países tiende a tener salarios más altos y cuál tiende a tener salarios más bajos y por qué?

B. De acuerdo al histograma: ¿qué rango salarial es más popular por país?

ii. Escriba otras conclusiones que pueda extraer de las estadísticas y los gráficos.

(d) **Código:** Considere tres tipos de salario anual: **bajo** ($\leq 10,000$), **medio** ($> 10,000$ y $\leq 50,000$) y **alto** ($> 50,000$), transforme el salario anual de cada usuario a estas categorías. Solo para los países previamente seleccionados.

(e) **Código:** Usando el salario transformado del subpunto previo, hacer lo siguiente para los países seleccionados:

i. Haga un gráfico de barras de las frecuencias de personas en cada categoría salarial por país.

ii. Calcule la probabilidad condicional para determinar qué país tiene una mayor probabilidad de tener un salario alto ($> 50,000$), y cuál tiene mayor probabilidad de tener un salario bajo ($\leq 10,000$).

iii. Calcule el chi-cuadrado de Pearson para determinar si existe una relación entre el país y el salario anual.

(f) **Reporte:**

i. Copie los gráficos y las estadísticas calculadas del subpunto previo y organícelos para poder comparar los datos fácilmente.

ii. De acuerdo a los gráficos de barras, las probabilidades condicionales, y las pruebas de chi-cuadrado de Pearson calculadas previamente para los países seleccionados, explique lo siguiente:

A. ¿Cuál es la categoría salarial más frecuente por país?

B. ¿Qué país tiene mayor probabilidad de tener un salario alto ($> 50,000$) y cuál tiene mayor probabilidad de tener un salario bajo ($\leq 10,000$), y por qué?

C. Si existe o no relación entre el país y el salario anual, ¿por qué?.

2. **Análisis por tipo de trabajo por país.** Considerando los datos de los países seleccionados de los puntos previos, haga lo siguiente y explique los subpuntos correspondientes.

(a) **Código:** De acuerdo a cada país, haga un gráfico de barras considerando solo los 3 trabajos más populares.

(b) **Código:** Calcule el resumen de los cinco números (mínimo, primer cuartil, mediana, tercer cuartil y máximo), la media, y la desviación estándar para el salario

anual por tipo de trabajo por país. Haga los diagramas de cajas e histogramas (con 10 bins) correspondientes.

(c) **Reporte:**

i. Copie los gráficos y las estadísticas previamente calculadas por tipo de trabajo y país. Organice las estadísticas y los gráficos para poder comparar fácilmente los datos entre tipos de trabajos y países. De acuerdo a las estadísticas, y el número de personas por tipo de trabajo por país:

A. ¿Qué tipo de trabajo tiende a tener salarios más altos en cada país y cuál tiende a tener salarios más bajos y por qué?

B. ¿Qué tipo de trabajo tiende a tener salarios más altos en general (entre los países seleccionados) y por qué?

C. De acuerdo al histograma: ¿qué rango salarial es más popular por tipo de trabajo en cada país?

ii. Escriba otras conclusiones que pueda extraer de las estadísticas y los gráficos.

(d) **Código:** Considere tres tipos de salario anual: **bajo** ($\leq 10,000$), **medio** ($> 10,000$ y $\leq 50,000$) y **alto** ($> 50,000$), transforme el salario anual de cada usuario a estas categorías. Solo para los países previamente seleccionados.

(e) **Código:** Usando el salario transformado del subpunto previo, hacer lo siguiente para los países seleccionados:

i. Haga un gráfico de barras de las frecuencias de las personas en cada categoría salarial por tipo de trabajo en cada país. Considere únicamente los 3 trabajos más populares en cada país.

ii. Calcule la probabilidad condicional para determinar qué tipo de trabajo tiene una mayor probabilidad de tener un salario alto ($> 50,000$) en cada país, y cuál tiene una mayor probabilidad de tener un salario bajo ($\leq 10,000$) en cada país.

iii. Calcule el chi-cuadrado de Pearson para determinar si existe una relación entre el tipo de trabajo y el salario anual en cada país.

(f) **Reporte:**

i. Copie los gráficos y las estadísticas previamente calculadas y organícelas de forma que los datos se puedan comparar fácilmente.

ii. De acuerdo a los gráficos de barras, las probabilidades condicionales, y las pruebas de chi-cuadrado de Pearson calculadas previamente para los países seleccionados, explique los siguiente:

A. ¿Cuál es el salario más frecuente por tipo de trabajo en cada país?

B. ¿Qué tipo de trabajo tiene una mayor probabilidad de tener un salario alto ($> 50,000$) en cada país, y cuál tiene una mayor probabilidad de tener un salario bajo ($\leq 10,000$) en cada país, y por qué?

C. Si existe o no relación entre el tipo de trabajo y el salario anual en cada país, ¿por qué?.

3. Considerando los datos de los países seleccionados en los puntos anteriores, haga lo siguiente y explique los subpuntos correspondientes.

   (a) **Código:** Calcule la correlación de Pearson entre **YearsCodePro** y el salario anual, por país.

   (b) **Reporte:** Copie las estadísticas obtenidas y organícelas para poder comparar fácilmente los datos. Explique si hay o no correlación entre las variables previas en cada país.

   (c) **Código:** Considere dos niveles de educación: **básico** ($<$ Bachelor's degree) y **alto** ($\geq$ Bachelor's degree), y tres tipos de salario anual **bajo** ($\leq 10,000$), **medio** ($> 10,000$ y $\leq 50,000$) y **alto** ($> 50,000$), transforme el nivel educacional y el salario anual de cada usuario en estas categorías. Únicamente para los 4 países seleccionados.

   (d) **Código:** Usando los datos transformados previamente, hacer lo siguiente para cada país seleccionado:

      i. Haga un gráfico de barras de las frecuencias de cada categoría de nivel educativo por país.

      ii. Calcule el chi-cuadrado de Pearson y la probabilidad condicional para determinar lo siguiente:

         A. Si existe una relación entre el nivel educacional y el salario anual.

         B. Cuál de los dos niveles educativos tiene una mayor probabilidad de tener un salario alto, y cuál tiene una mayor probabilidad de tener un salario bajo.

   (e) **Reporte:**

      i. De acuerdo a la prueba del chi-cuadrado de Pearson y la probabilidad condicional previamente calculadas, explique lo siguiente:

         A. Si existe una relación entre el nivel educacional y el salario anual, ¿por qué?.

         B. ¿Cuál de los dos niveles educativos tiene una mayor probabilidad de tener un salario alto ($> 50,000$) y cuál tiene la mayor probabilidad de un salario bajo ($\leq 10,000$), y por qué?

   (f) **Código:** Para cada país seleccionado, haga un gráfico de barras de los 3 lenguajes más populares.

   (g) **Código:** Calcule el resumen de los cinco números (mínimo, primer cuartil, mediana, tercer cuartil y máximo), la media y la desviación estándar para el salario anual por lenguaje por país. Haga los diagramas e histogramas (con 10 bins) correspondientes.

   (h) **Reporte:**

      i. Copie los gráficos y las estadísticas previamente calculadas por lenguaje y país. Organice las estadísticas y los gráficos para poder comparar fácilmente los datos entre lenguajes y países. De acuerdo a las estadísticas, y el número de personas por lenguaje por país:

A. ¿Qué lenguaje tiende a tener salarios más altos en cada país y cuál tiende a tener salarios más bajos y por qué?

B. ¿Qué lenguaje tiende a tener salarios más altos en general (entre los países seleccionados) y por qué?

C. De acuerdo al histograma: ¿qué rango salarial es más popular por lenguaje en cada país?

ii. Escriba otras conclusiones que pueda extraer de las estadísticas y los gráficos.

4 Extra

De acuerdo con la clase del día **jueves 20 de marzo**, la implementación en código para la prueba del chi-cuadrado de Pearson será individual. **PROHIBIDO** usar alguna librería que haga el cálculo directamente. Para obtener el valor de referencia o valor crítico, se puede usar la librería SciPy:

```
from scipy.stats import chi2
var_ref = chi2.isf(α, df)
```

o la Tabla: **Upper-tail critical values of chi-square distribution with ν degrees of freedom** disponible en:
https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm