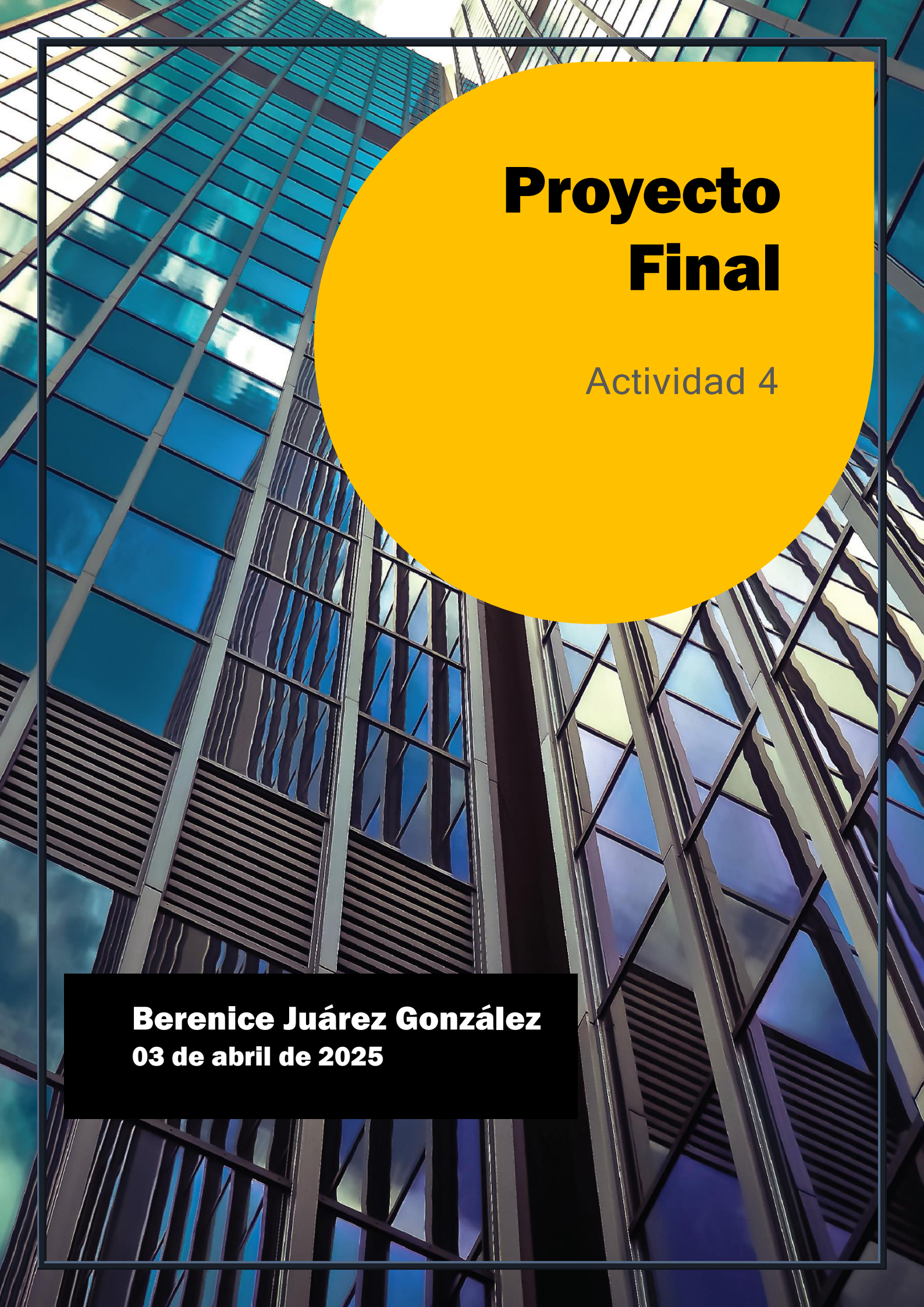


Proyecto Final

Actividad 4

Berenice Juárez González

03 de abril de 2025

INTRODUCCIÓN

En esta actividad, se ha realizado un conjunto de datos sobre los salarios de mujeres, identificando patrones, problemas y realizando visualizaciones.

CARGA Y EXPLORACIÓN DEL DATASET

- Se cargó el dataset desde un archivo CSV. Localizado en una ruta específica.

```
> print("2. Cargando el dataset desde un archivo CSV...")
[1] "2. Cargando el dataset desde un archivo CSV..."
> ruta <- "C:/Users/bjuar/OneDrive/Documentos/WomenIn/salarios_mujeres.csv"
> datos_a <- read_csv(ruta)
Rows: 50000 Columns: 3
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): Genero
dbl (2): Edad, Salario
```

- Se revisaron las primeras filas para conocer su tamaño total.

```
> print("3. Explorando el dataset cargado...")
[1] "3. Explorando el dataset cargado..."
>
> #ver las primeras filas
> print("Primeras 6 filas del dataset:")
[1] "Primeras 6 filas del dataset:"
> head(datos_a)
# A tibble: 6 × 3
   Edad Salario Genero
  <dbl>   <dbl> <chr>
1    53   23652 Mujer
2    43    6137 Mujer
3    29    5740 Mujer
4    57   23652 Mujer
5    22    4090 Mujer
6    35    6137 Mujer
```

- Se obtuvieron las dimensiones del dataset (filas y columnas).

```
> print("Información general del dataset:")
[1] "Información general del dataset:"
> str(datos_a)
spc_tbl_ [50,000 × 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Edad   : num [1:50000] 53 43 29 57 22 35 53 33 37 25 ...
 $ Salario: num [1:50000] 23652 6137 5740 23652 4090 ...
 $ Genero : chr [1:50000] "Mujer" "Mujer" "Mujer" "Mujer" ...
 - attr(*, "spec")=
  .. cols(
  ..   Edad = col_double(),
  ..   Salario = col_double(),
  ..   Genero = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
>
> #dimension del dataset
> print("Dimensiones del dataset (filas x columnas):")
[1] "Dimensiones del dataset (filas x columnas):"
> dim(datos_a)
[1] 50000     3
```

- Se listaron los nombres de las columnas para identificar las variables disponibles.

```
> #nombres de las columnas
> print("Nombres de las columnas:")
[1] "Nombres de las columnas:"
> names(datos_a)
[1] "Edad"    "Salario" "Genero"
>
```

- Se generó un resumen estadístico que incluyó medidas como la media, mediana y los valores máximos y mínimos de las variables numéricas.

```
> #resumen estadistico del dataset
> print("Resumen estadístico del dataset:")
[1] "Resumen estadístico del dataset:"
> summary(datos_a)
      Edad          Salario         Genero
 Min.   :15.00   Min.   : 4090   Length:50000
 1st Qu.:27.00   1st Qu.: 5740   Class :character
 Median :39.00   Median : 6137   Mode  :character
 Mean   :39.48   Mean   :12141
 3rd Qu.:52.00   3rd Qu.:23652
 Max.   :64.00   Max.   :24988
                 NA's   :2478
```

IDENTIFICACIÓN DE PROBLEMAS EN LOS DATOS

Valores Nulos: Se encontraron un total de 2478 valores nulos distribuidos de la siguiente manera:

```
                 NA's   :2478
> #Identificacion de problemas en los datos
> print("4. Identificando problemas en los datos...")
[1] "4. Identificando problemas en los datos..."
>
> #Contar valores nulos
> total_nulos <- sum(is.na(datos_a))
> print(paste("Total de valores nulos en el dataset:", total_nulos))
[1] "Total de valores nulos en el dataset: 2478"
>
> # Valores nulos por columna
> print("Número de valores NA por columna:")
[1] "Número de valores NA por columna:"
> print(colSums(is.na(datos_a)))
   Edad Salario  Genero
      0    2478       0
>
```

Registros Duplicados: Se identificaron 49900 registros duplicados, los cuales pueden generar problemas en el análisis y deben ser eliminados o revisados.

```
> # Identificar registros duplicados
> duplicados <- datos_a %>% filter(duplicated(.))
> print(paste("Número de registros duplicados:", nrow(duplicados)))
[1] "Número de registros duplicados: 49900"
```

CÁLCULO DE ESTADÍSTICAS DESCRIPTIVAS

Media de Edad: 39.4786 años, lo que indica el promedio de edad de las personas en el dataset.

```
> # Calcular la media de la edad
> media <- datos_a %>% summarise(Media_Edad = mean(Edad, na.rm = TRUE))
> print(paste("Media de Edad:", media$Media_Edad))
[1] "Media de Edad: 39.4786"
```

Mediana de Edad: 39 años, valor que representa la edad central cuando los datos están ordenados.

```
> # Calcular la mediana de la edad
> mediana <- datos_a %>% summarise(Mediana_Edad = median(Edad, na.rm = TRUE))
> print(paste("Mediana de Edad:", mediana$Mediana_Edad))
[1] "Mediana de Edad: 39"
```

Moda del Salario: 6137, que es el valor salarial que mas se repite dentro del conjunto de datos.

```
> # Calcular la moda del salario
> moda <- datos_a %>%
+   count(Salario) %>%
+   filter(n == max(n)) %>%
+   pull(Salario)
> print(paste("Moda de Salario:", moda))
[1] "Moda de Salario: 6137"
```

FILTRADO DE DATOS

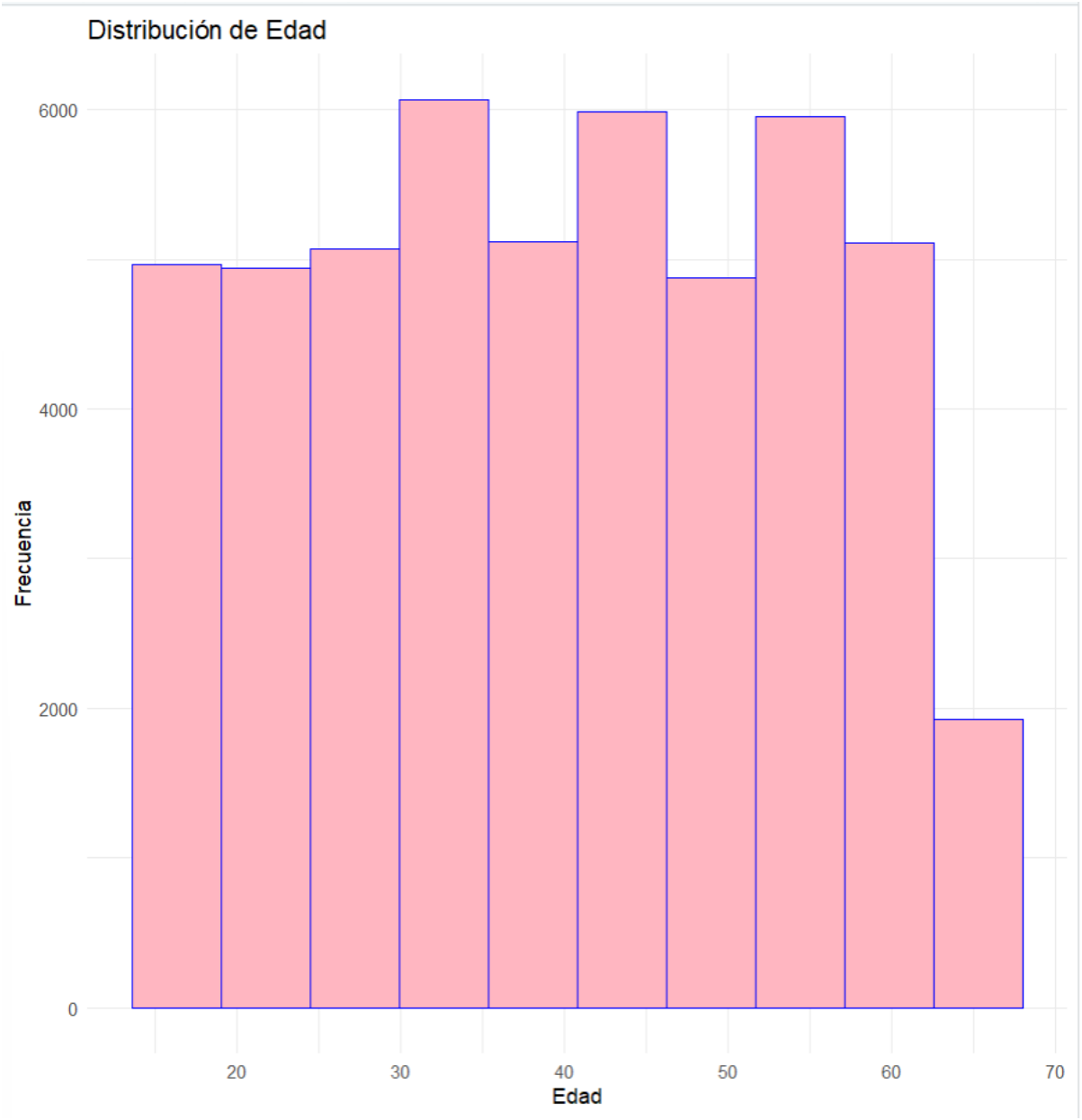
Para realizar un análisis específico, se filtraron los datos considerando exclusivamente registros de mujeres con edades de 25 y 30 años. Esto permite observar tendencias salariales específicas en este grupo de edad.

.

```
> # Filtrado de datos
> print("6. Filtrando datos por edades de 25 y 30 años...")
[1] "6. Filtrando datos por edades de 25 y 30 años..."
>
> datos_filtrados <- datos_a %>% filter(Edad %in% c(25, 30))
>
> print("Datos filtrados con edades de 25 y 30 años:")
[1] "Datos filtrados con edades de 25 y 30 años:"
> head(datos_filtrados)
# A tibble: 6 × 3
   Edad Salario Genero
  <dbl>   <dbl> <chr>
1    25    5740 Mujer
2    25    5740 Mujer
3    30    5740 Mujer
4    30    5740 Mujer
5    25    5740 Mujer
6    25    5740 Mujer
```

VISUALIZACIÓN DE DATOS

Se generó un **histograma** de la variable "Edad" para analizar la distribución etaria en el dataset. La gráfica muestra la frecuencia de aparición de diferentes edades en la base de datos, permitiendo detectar posibles sesgos o concentraciones de datos en ciertos rangos de edad.



CONCLUSIONES

- Se identificaron problemas como valores nulos y registros duplicados que pueden afectar la calidad de los datos y deben ser tratados adecuadamente.
- Las medidas de tendencia central (media, mediana y moda) permitieron comprender mejor la distribución de las variables Edad y Salario.
- La distribución de edades se visualizó a través de un histograma, lo que ayudó a identificar patrones en los datos.