Tarea 1

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(crayon)
```

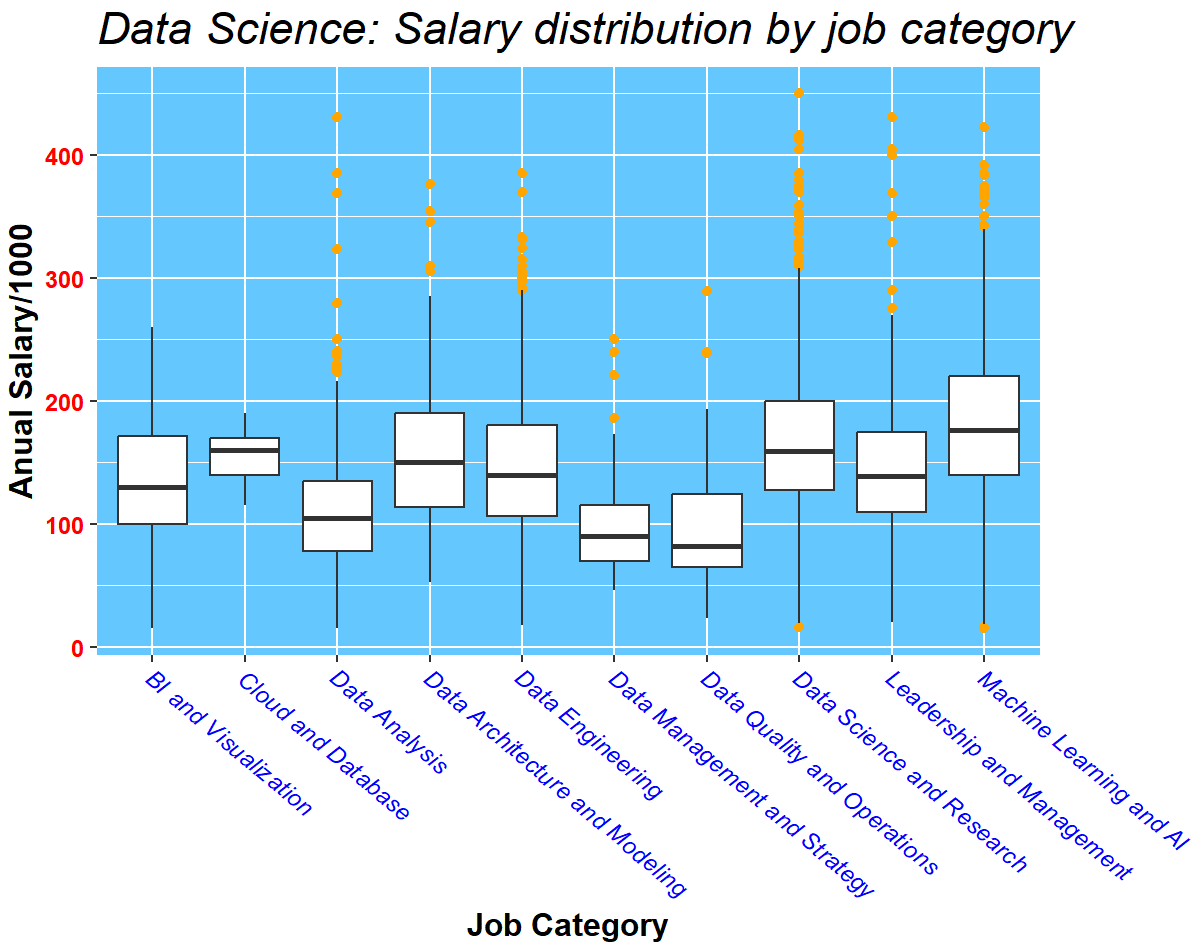
```
##
## Attaching package: 'crayon'
```

```
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

```
df <- read.csv(file = "jobs_in_data.csv")
```

Distribución de salario anual por categoría de trabajo

```
boxPlot <- ggplot(df, aes(x = job_category, y = salary_in_usd/1000)) +
  geom_boxplot(outlier.color = "orange") +
  theme(axis.text.y = element_text
        (size = 9,
          color = "red",
          face = "bold"),
        axis.text.x = element_text
        (size = 9,
          color = "blue",
          face = "italic",
          angle = 320,
          vjust = 0.9,
          hjust = -0.009),
        title = element_text
        (size = 14,
          color = "Black",
          face = "italic"),
        axis.title = element_text
        (size = 12,
          color = "Black",
          face = "bold"),
        panel.background = element_rect(fill = "#67c9ff")
        ) +
  labs(x = "Job Category",
       y = "Anual Salary/1000",
       title = "Data Science: Salary distribution by job category"
       )
boxPlot + coord_fixed(ratio = (6/11)/41)
```



- **A. Cuál es la categoría de trabajo donde el salario es más uniforme ? Explique porqué.**

  El trabajo que representa este comportamiento es en el campo del "machine Learning and AI", la razón principal es porque presenta una linea más constante que la demás, dando así un mayor abanico salarial.

- **B. Cúal es la categoría de trabajo donde el salario varía más ? Explique porqué.**

  El trabajo que muestra esta anomalia es en "Data science and research. La razón principalmente es porque se muestra muchos casos atípicos(Es decir, eventos que se devía de lo común) ya que se sobresalen muchos puntos de la barra de la caja, por lo que a partir de este hecho se puede decir que sus salarios varian bastante con respecto a esta área.

- **C. Cúal es la categoría de trabajo en donde se gana más salario en promedio ? Explique porqué.**

  Un campo en donde se gana más según el promedio es en "machine Learning and AI", ya que su segundo cuartil(Q2) está por encima de los demás. Esto quiere decir que es muy apreciado en el mercado de esta área, por lo que su media salarial es la mayor de todas. Es esperable que al ejercer en este espacio lo usal es obtener una remuneración de esa cantidad.

- **D. Cuál es la categoría de trabajo en donde se gana menos salario en promedio ? Explique porqué.**

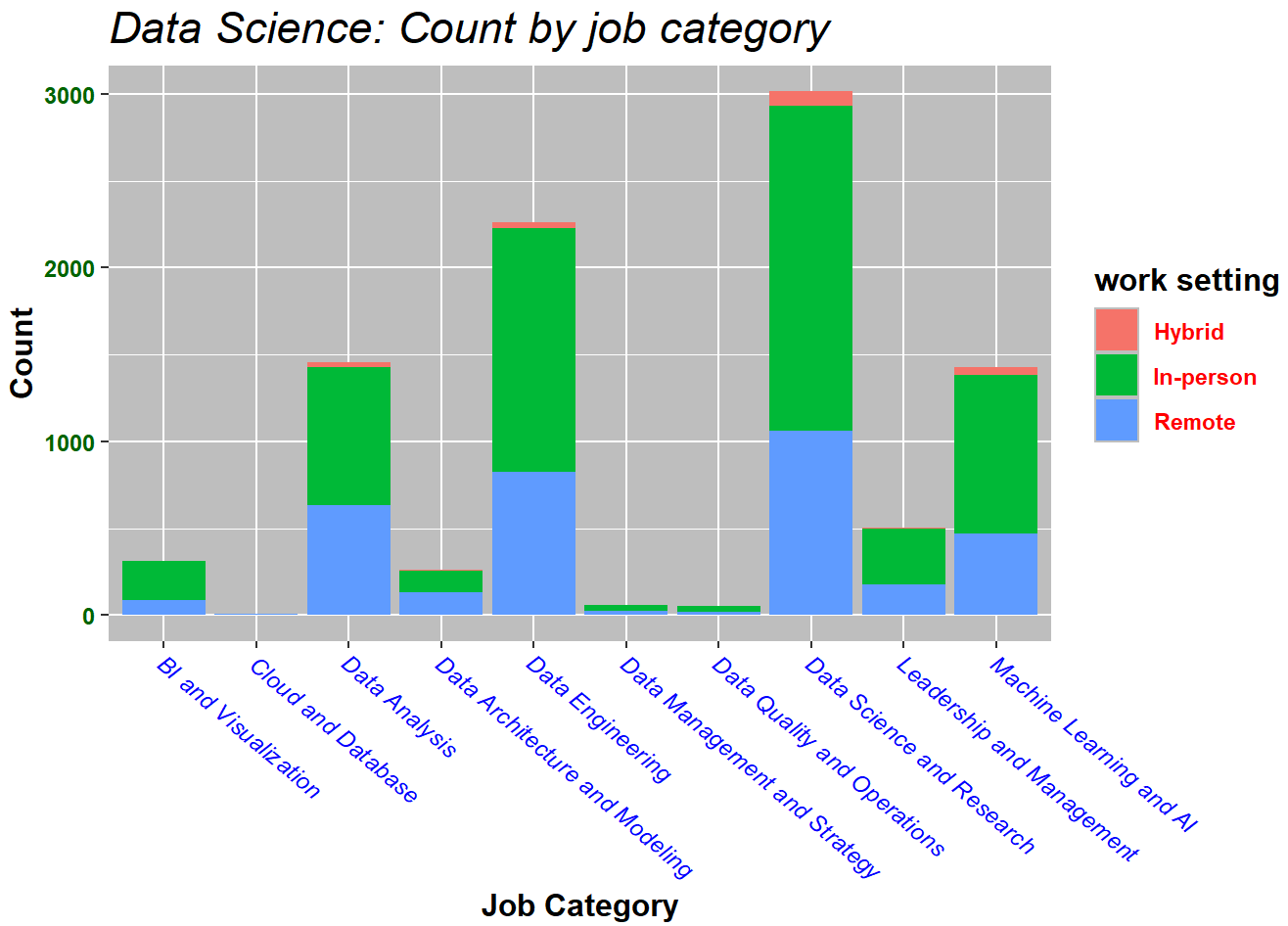
  "Data Quality and Operations" es la categoría de la cúal se obtiene una menor remuneración en promedio a comparación de las demás. Este hecho se presenta ya que su Q2 es la menor de todas

Cantidad de trabajos por categoría

```
df <- read.csv("jobs_in_data.csv")

ejeX <- aes(x=job_category, fill=work_setting)

ggplot(df, ejeX ) +
  geom_bar() +
  theme(axis.text.y = element_text
        (size = 9,
          color = "darkgreen",
          face = "bold"),
        axis.text.x = element_text
        (size = 9,
          color = "blue",
          face = "italic",
          angle = 320,
          vjust = 0.9,
          hjust = -0.009),
        title = element_text
        (size = 14,
          color = "Black",
          face = "italic"),
        axis.title = element_text
        (size = 12,
          color = "Black",
          face = "bold"),
        panel.background = element_rect(fill = "gray"),
        legend.title = element_text
        (size = 12,
          color = "Black",
          face = "bold")
        ) +
  labs(x = "Job Category",
       y = "Count",
       title = "Data Science: Count by job category",
       fill = "work setting"
       ) +
  theme(legend.text = element_text
        (
          color = "red",
          face = "bold"),
        axis.title = element_text(size = 12)
        )
```



- **A. Cuál es la categoría de trabajo donde existe mayor cantidad de puestos y es mejor pagado ? Explique porqué.**

  La categoría que le corresponde este comportamiento es en "Data Science and Research". La razón de esto es por que en las barras muestra no solo que es superior en la cantidad de puestos en Remoto, sino que también el resto de aspectos(In-person y Hybrid), por lo que ese sería el puesto que uno puede optar más facilmente si uno tiene los conocimientos necesarios, encima tiene unos salarios muy competitivos y su alta demanda le otorga este puesto.

- **B. Cúal es la categoría de trabajo donde existe un buen salario pero hay pocos puestos disponibles ? Explique porqué.**

  La categoría de trabajo que coincide con esos parametros es en "Cloud and Database", este a pesar de ser el que menos puestos tiene, a la vez consigue unos salarios buenos a comparación de otros que tienen más puestos. Esto se logra viendo en el grafico de cajas su Q2, mientras que en el grafico de barras apenas se aprecia.

- **C. Cúal es la categoría de trabajo peor pagado y con menor cantidad de puestos disponibles ? Explique porqué.**

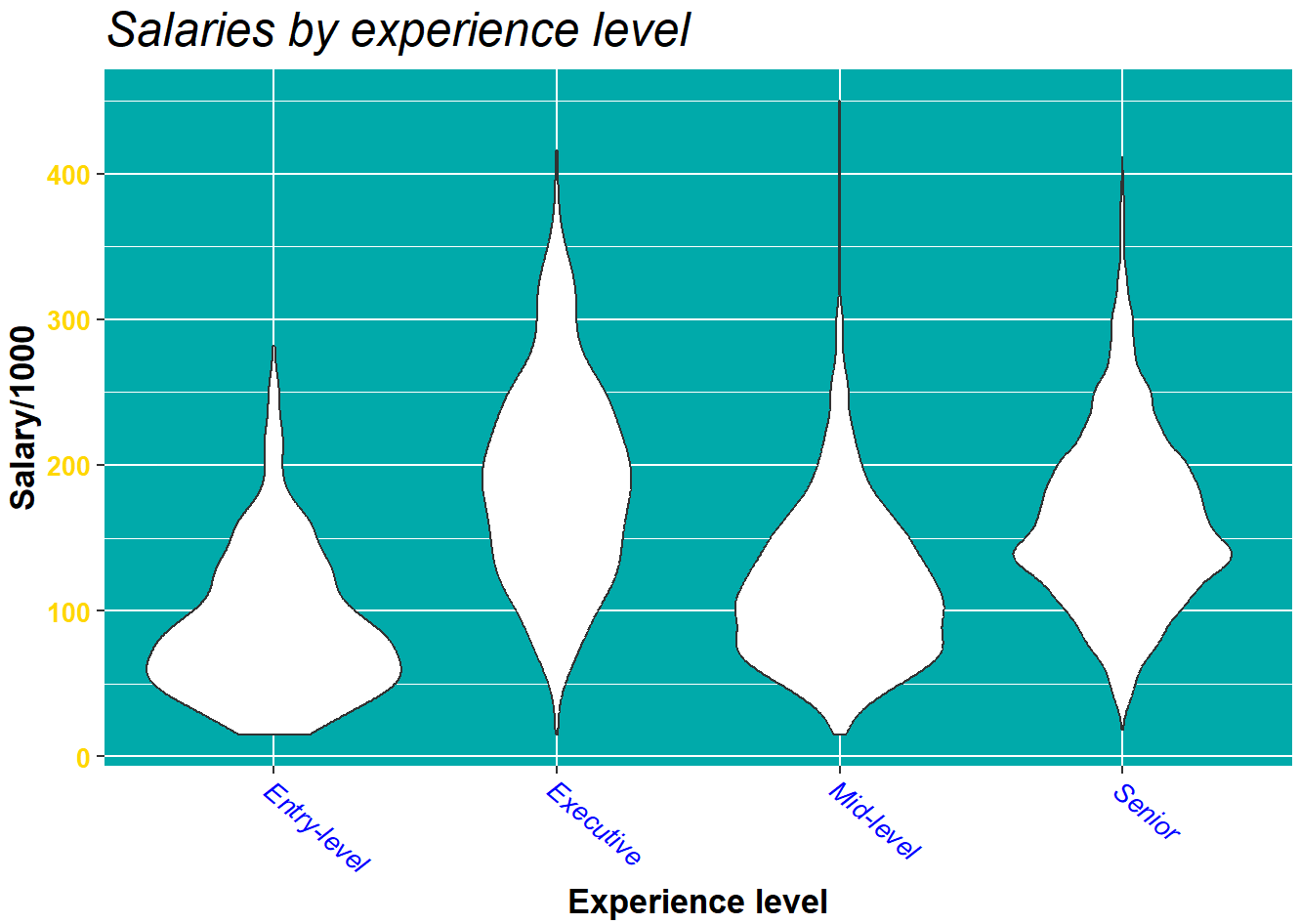
  La categoria que le corresponde esa información se llama "Data Quality and Operations". Esto se aprecio al lograr ver que cuenta con pocos puestos en el grafico de barras, a su vez su media salarial es la más baja de todas.

- **D. Cuál es la categoría de trabajo con mayor cantidad de puestos pero salarios bajos ? Explique porqué.**

  "Data analisis" es el trabajo que coincide con dicha propiedad. Esto se logra encontrar gracias al grafico de barras que este una cantidad bastante buena en el sector TI, pero sus salarios son de los más bajos conciderando su buena salida en el mercado, esto último se logra observar por el grafico de cajas.

Salarios por nivel de experiencia

```
ggplot(df, aes(x = experience_level, y = salary_in_usd/1000)) +
  geom_violin() +
  theme(axis.text.y = element_text
        (size = 10,
          color = "gold",
          face = "bold"),
        axis.text.x = element_text
        (size = 10,
          color = "blue",
          face = "italic",
          angle = 320,
          vjust = 0.9,
          hjust = -0.009),
        title = element_text
        (size = 15,
          color = "Black",
          face = "italic"),
        axis.title = element_text
        (size = 13,
          color = "Black",
          face = "bold"),
        panel.background = element_rect(fill = "#00AAAA")
        ) +
  labs(x = "Experience level",
       y = "Salary/1000",
       title = "Salaries by experience level")
```



- **A. En promedio cúal es el nivel de experiencia con los peores salarios ? Explique porqué.**

  El nivel de experiencia que toma en cuenta esos parametros es el de "Entry-level. Se muestra muy bien en el gráfico que el grueso de la sección es mucho más pronunciada en la parte baja, dando de está manera los indicios de que la mayoría obtiene esos salarios.

- **B. En promedio cuál es el nivel de experiencia con los mejores salarios ? Explique porqué.**

  El nivel de senior es el que cuenta con una redistribuición a comparación con el resto. Esto se observa en el ancho del apartado de Senior en la gráfica.

- **C. Cuál es el nivel de experiencia que tiene el salario más alto ? Explique porqué.**

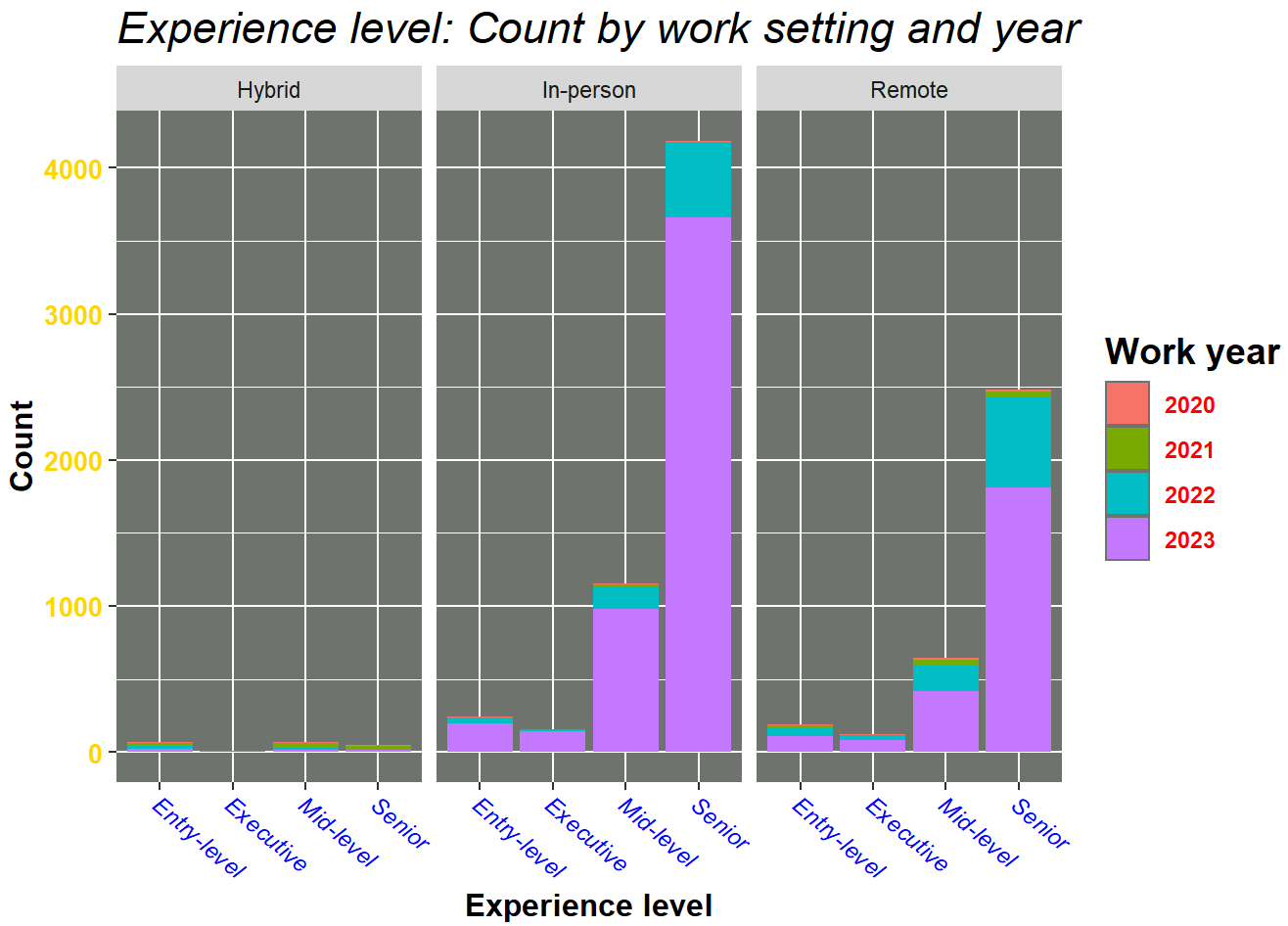
  Uno en primera instancia pensaría en el "mid-level", pero eso se podría considerar un caso atípico, realmente el nivel que tiene el salario más alto es el "Executive". Esto se logra determinar que en los niveles más altos que tiene una proporción adecuada para el puesto es este.

- **D. Cuál es el nivel de experiencia que tiene un rango de salarios más amplio ? Explique porqué.**

  El senior se muestra con un nivel bastante uniforme al resto, por lo que este se considera como el rango salarial más amplio.

Cantidad de puestos por nivel de experiencia y año

```
ggplot(df, aes(x = experience_level, fill = work_year)) +
  geom_bar(stat="count", aes(fill = as.factor(work_year))) +
  facet_wrap(~work_setting, nrow=1) +
  theme(axis.text.y = element_text
        (size = 10,
          color = "gold",
          face = "bold"),
        axis.text.x = element_text
        (size = 9,
          color = "blue",
          face = "italic",
          angle = 320,
          vjust = 0.9,
          hjust = -0.009),
        title = element_text
        (size = 14,
          color = "Black",
          face = "italic"),
        axis.title = element_text
        (size = 12,
          color = "Black",
          face = "bold"),
        panel.background = element_rect(fill = "#707770")
        ) +
  labs(x = "Experience level",
       y = "Count",
       title = "Experience level: Count by work setting and year",
       fill = "Work year") +
  theme(legend.text = element_text
        (
          color = "red",
          face = "bold"),
        axis.title = element_text(size = 12),
        legend.title = element_text(color = "black",
                                    face = "bold")
        )
```



- **A. Cuál es el año donde más se crearon trabajos remotos para personas con nivel senior? Explique porqué.**

  El año 2023 ha sido la época en donde más trabajos remotos se ha absorbido por parte de los senior. Esto se sabe gracias a que la barra es mucho más prominente a comparación del resto de años.

- **B. Qué nivel de experiencia tenía más posibilidad de obtener un puesto remoto en el 2023? Explique porqué.**

  El senior ha sido el candidato predilecto por el que tuvo más posibilidades de obtener un puesto en el año 2023. Esto se observa en la grafica como el trabajo remoto predomina en la sección del senior.

- **C. Para cúal nivel de experience se crearon, a través de los años, una mayor proporción de trabajos ? Explique porqué.**

  El predominante a lo largo de los años ha sido la posición de "Senior" como se puede observar en el gráfico, a excepción del año 2021 y 2020, del cúal los "mid-level" tuvierón un poco más representación en esos años.

- **D. Cuál fue el año que solo se ofrecieron trabajos presenciales para ejecutivos ? Explique porqué.**

  En el gráfico no se muestra dichos parametros de los cuales se pueda adjudicar a los ejecutivos. En cambio, un año en el que los ejecutivos hayan tenido la mayoría de propuesta en presencial ha sido el año 2023. Esto se logra al distingir los conteos que se puede determinar analisando la grafica de barras.